Directorate for Education
and Human Resources

Division of Research,
Evaluation and Dissemination

User-Friendly Handbook for Project Evaluation:

Science, Mathematics, Engineering and Technology Education

Co-Authors:

Floraline Stevens

Frances Lawrenz

Laure Sharp

Edited by:

Joy Frechtling

The Foundation provides awards for research in the sciences and engineering. The awardee is wholly responsible for the conduct of such research and preparation of the results for publication. The Foundation, therefore, does not assume responsibility for the research findings or their interpretation.

The Foundation welcomes proposals from all qualified scientists and engineers, and strongly encourages women, minorities, and persons with disabilities to compete fully in any of the research and related programs described here.

In accordance with federal statutes, regulations, and NSF policies, no person on grounds of race, color, age, sex, national origin, or disability shall be excluded from participation in, denied the benefits of, or be subject to discrimination under any program or activity receiving financial assistance from the National Science Foundation.

Facilitation Awards for Scientists and Engineers with Disabilities (FASED) provide funding for special assistance or equipment to enable persons with disabilities (investigators and other staff, including student research assistants) to work on an NSF project. See the program announcement or contact the program coordinator at (703) 306-1633.

Privacy Act and Public Burden

Information requested on NSF application materials is solicited under the authority of the National Science Foundation Act of 1950, as amended. It will be used in connection with the selection of qualified proposals and may be used and disclosed to qualified reviewers and staff assistants as part of the review process and to other government agencies. See Systems of Records, NSF-50, "Principal Investigator/Proposal File and Associated Records," and NSF-51, "Reviewer/Proposals File and Associated Records," 56 Federal Register 54907 (Oct. 23, 1991). Submission of the information is voluntary. Failure to provide full and complete information, however, may reduce the possibility of your receiving an award.

The public reporting burden for this collection of information is estimated to average 120 hours per response, including the time for reviewing instructions. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Herman G. Fleming, Reports Clearance Officer, Division of CPO, NSF, Arlington, VA. 20330; and the Office of Management and Budget, Paperwork Reduction Project (3145-0058), Wash., D.C. 20503.

The National Science Foundation has TTD (Telephonic Device for the Deaf) capability, which enables individuals with hearing impairment to communicate with the Foundation about NSF programs, employment, or general information. This number is (703) 306-0090.

ACKNOWLEDGMENTS

Appreciation is expressed to the Project Directors, NSF staff, and Advisory Committee members who reviewed drafts of this document. Specifically we would like to thank Jack Bookman, Cha Guzman, Kathleen Martin, Mildred Murray-Ward, Lynette Padmore, and Sharon Reynolds who reviewed the Handbook from a project perspective; NSF staff members Susan Gross, Larry Enochs, Conrad Katzenmeyer, and Larry Suter who reviewed it from a Foundation perspective; and members of the Advisory Committee for Research, Evaluation, and Dissemination, especially Wayne Welch and Nick Smith, who brought an important outside perspective.

USER-FRIENDLY HANDBOOK FOR PROJECT EVALUATION: Science, Mathematics, Engineering and Technology Education

Table of Contents

ABOUT THE AUTHORS

This Handbook is the product of several authors. Originally, it began as two separate papers describing various aspects of the evaluation process. Sensing a need for a more complete document, the National Science Foundation asked that the papers be used as the basis for a Handbook specially tailored for projects supported by the Directorate for Education and Human Resource Development (EHR). The Handbook has grown a bit beyond the original papers and the papers have been somewhat reshaped, but the message and advice contained reflect the solid advice of the NSF staff that originated it. Credit for the Handbook goes to:

Dr. Floraline Stevens

Dr. Stevens is currently a Program Director for Evaluation in the Evaluation Section of the Division of Research, Evaluation, and Dissemination (RED). While at NSF she is on an Interagency Personnel Assignment from the Los Angeles Public Schools where she is Director of Research and Evaluation. Dr. Stevens conceived the idea for this Handbook and drafted the original versions of Chapters One, Four, and Six, and the Glossary.

Dr. Frances Lawrenz

Dr. Lawrenz is a Professor at the University of Minnesota and Director of Graduate Studies in the Department of Curriculum and Instruction. While at NSF she was an Evaluation Specialist in RED. She served as a advisor on evaluation to other divisions and provided support to the evaluation groups of the Federal Coordinating Committee on Science, Engineering, and Technology. Dr. Lawrenz wrote the original version of Chapter Two.

Ms. Laure Sharp

Ms. Laure Sharp is a social science researcher, who has published extensively in the areas of survey research, high school and post high school education, and social program evaluation. She was formerly a senior research associate and assistant director at the Bureau of Social Science Research in Washington, D.C. She is currently a consultant to Booz·Allen & Hamilton Inc. and is the author of Chapter Three.

Additional writing and editing was done by the staff of Booz·Allen & Hamilton Inc. Dr. Joy Frechtling refined the original chapters and provided overall technical editing.

INTRODUCTION: ABOUT THIS HANDBOOK

This Handbook was developed to provide Principal Investigators and Project Evaluators working with the National Science Foundation's Directorate for Education and Human Resource Development (EHR) with a basic understanding of selected approaches to evaluation. It is aimed at people who need to learn more about both what evaluation can do and how to do an evaluation, rather than those who already have a solid base of experience in the field. It builds on firmly established principles, blending technical knowledge and common sense to meet the special needs of NSF's programs and projects and those involved in them.

NSF supports a wide range of programs aimed at improving the status of mathematics, science, technology, and engineering in our schools and increasing the participation of students at every level of the educational system. Each program funds many projects, some of which are broad-based and systemic in nature; others which are more specifically focused on a part or a small number of parts of the educational system. Evaluation is important to each one of these.

NSF and the Principal Investigators themselves need to know what these projects and programs are accomplishing, and what it is about them that makes them work or may stand in the way of success. Although the approaches to evaluation selected may differ depending on the nature of the program or project, its goals, and where it is in its "life cycle," the Foundation firmly believes that each program and project can be improved by soundly conducted evaluation studies. While the information in this Handbook should be useful in evaluating programs as well as projects, it is primarily targeted at project evaluation, which may be conducted by a member of the project staff or by an outside evaluator.

The Handbook discusses quantitative and qualitative evaluation methods, but the emphasis is on quantitative techniques for conducting outcome evaluations, those designed to assess the results of NSF funded innovations and interventions. Although there is much interest in the evaluation community in a less traditional and more qualitative approach to evaluation, at the present time this approach seldom meets NSF requirements, especially for Summative Evaluations. For activities dependent on federal funding, which are

subject to periodic funding decisions by NSF managers as well as the Office of Management and Budget (OMB) and congressional staffs and decision-makers, emphasis is still on quantitatively measurable outcome information: did the treatment or innovation (the program funded by the federal agency) result in outcomes which can be attributed to these federal expenditures and might not have occurred without these expenditures? As stated in a recent report issued by GAO (the Government Accounting Office, which is the agency charged with oversight of government programs for the U.S. Congress):

"Over the next few years, the federal government will face powerful opposing pressures: the need on the one hand to reduce the federal deficit, and the demand, on the other, for a federal response to some potentially expensive domestic problems. . . (The need is for) program effectiveness evaluations which estimate the effects of federal programs using statistical analysis of outcomes (such as educational achievement test scores or conditions of housing) for groups of persons receiving program services compared with similar groups of non participants."

Obviously, decision makers at the highest levels of the executive and legislative branches of government are looking for traditional "effectiveness indicators" although, as we have emphasized in the Handbook, these are often very difficult for evaluators to establish.

To develop this Handbook we have drawn on the skills of both NSF staff familiar with the Foundation's educational programs and outside evaluators who have experienced the challenge of examining projects in a real-world setting. This Handbook is not intended to be a theoretical treatise, but rather a practical guide to evaluating NSF/EHR funded projects.

The Handbook addresses several topics. The first four chapters focus on designing and implementing evaluation studies:

- Chapter One describes the various types of evaluation prototypes
- Chapter Two presents an overview of the evaluation process

- Chapter Three describes the collection and analysis of data
- Chapter Four examines report writing.

The remaining chapters provide support materials:

- Chapter Five contains examples of project evaluations, illustrating the prototypes described earlier, and highlighting strengths and weaknesses of the approaches described.
- Chapter Six provides information on selecting an evaluator.
- Chapter Seven provides a glossary.
- Finally (for those who are challenged or intrigued by what they have read here, or feel the need to learn more), Chapter Eight provides supplemental references in the form of an annotated bibliography.

In addition to evaluation, Project Directors need to plan the dissemination of project outcomes to a broader audience. A separate publication dealing with dissemination guidelines has been prepared by NSF.

REFERENCE

Government Accounting Office (1992). *Program Evaluation Issues.* GAO/OCG-93-6TR.

CHAPTER ONE: EVALUATION PROTOTYPES

The purpose of this chapter is to help Principal Investigators and Project Evaluators think practically about evaluation and the kinds of information evaluations can provide. We start with the assumption that the term "evaluation" describes different models or prototypes that suit different purposes at different stages in the life of a project. A major goal of this chapter is to help Principal Investigators and Project Evaluators understand what some of these different prototypes are and to assist them in using different approaches to evaluation to meet these varying needs.

What is Evaluation?

The notion of evaluation has been around a long time—in fact, the Chinese had a large functional evaluation system in place for their civil servants as long ago as 2000 B.C. Not only does the idea of evaluation have a long history, but it also has varied definitions. Evaluation means different things to different people and takes place in different contexts. Thus, evaluation can be synonymous with tests, descriptions, documentation, or management. Many definitions have been developed, but a comprehensive definition is presented by the Joint Committee on Standards for Educational Evaluation (1981):

Evaluation means different things to different people and takes place in different contexts.

"Systematic investigation of the worth or merit of an object. . ."

This definition centers on the goal of using evaluation for a purpose. Evaluations should be conducted for action-related reasons, and the information provided should facilitate deciding a course of action.

Over the years evaluation has frequently been viewed as an adversarial process. Its main use has been to provide "thumbs-up" or "thumbs down" about a program or project. In this role, it has all too often been considered by program or project Directors as an external imposition which is threatening, disruptive, and not very helpful to Project staff. Our contention is that while this may be true in some situations, this is not the case in all, nor even in most, evaluation efforts. And, today in contrast to a decade or two ago, the view is gaining ground that evaluation should be a tool that not only measures, but can contribute to, success.

CHAPTER THREE: DESIGN, DATA COLLECTION AND DATA ANALYSIS

In Chapter Two we outlined the steps in the development and implementation of an evaluation. Another name for that chapter could be "the Soup to Nuts" of evaluation because of its broad-based coverage of issues. In this chapter we focus more closely on selected technical issues, the "Nuts and Bolts" of evaluation, issues that generally fall into the categories of design, data collection and analysis.

In selecting these technical issues, we were guided by two priorities:

> We devoted most attention to topics relevant to quantitative evaluations, because, as emphasized in the introduction, in order to be responsive to executive and congressional decisionmakers, NSF is usually required to furnish outcome information based on quantitative measurement.
>
> We have given the most extensive coverage to topics for which we have located few concise reference materials suitable for NSF/EHR project evaluators. But for all topics, we urge project staff who plan to undertake comprehensive evaluations to make use of the reference materials mentioned in this chapter and in the annotated bibliography.

The chapter is organized into four sections:

- How do you design an evaluation?
- How do you choose a specific data collection technique?
- What are some major concerns when collecting data?
- How do you analyze the data you have collected?

How Do You Design an Evaluation?

Once you have decided the goals for your study and the questions you want to address, it is time to design the study. What does this mean? According to Scriven (1991) design means:

"The process of stipulating the investigatory procedures to be followed in doing a certain evaluation."

Thoughtful analysis, sensitivity, common sense, and creativity are all needed to make sure that the actual evaluation provides information that is useful and credible.

Designing an evaluation is one of those "good news — bad news" stories. The good news is that there are many different ways to develop a good design. The bad news is that there are many ways to develop bad designs. There is no formula or simple algorithm that can be relied upon in moving from questions to an actual study design. Thoughtful analysis, sensitivity, common sense, and creativity are all needed to make sure that the actual evaluation provides information that is useful and credible.

This section examines some issues to consider in developing designs that are both useful and methodologically sound. They are:

- Choosing an approach
- Selecting a sample
- Deciding how many times to measure

Choosing an Approach

Since there are no hard and fast rules about designing the study, how should the evaluator go about choosing the procedures to be followed? This is usually a 2-step process. In step 1, the evaluator makes a judgment about the main purpose of the evaluation, and about the over-all approach which will provide the best framework for this purpose. This judgment will lead to a decision whether the methodology will be essentially qualitative (relying on case studies, observations, and descriptive materials) or whether the method should rely on statistical analyses, or whether a combined approach would be best. Will control or comparison groups be part of the design? If so, how should these groups be selected?

While some evaluation experts feel that qualitative evaluations should not be treated as a technical, scientific process (Guba and Lincoln, 1989) others (for example, Yin, 1989) have adopted design strategies which satisfy rigorous scientific requirements. Conversely, competently executed quantitative studies will have qualitative components. The experienced evaluator will want to see a project in action and

conduct observations and informational interviews before designing instruments for quantitative evaluation; he or she will also provide opportunities for "open-ended" responses and comments during data collection.

There is a useful discussion about choosing the general evaluation approach in Herman, Morris, and Fitz-Gibbon (1987) which concludes with the following observation:

"There is no single correct approach to all evaluation problems. The message is this: some will need a quantitative approach; some will need a qualitative approach; probably most will benefit from a combination of the two."

In all cases, once fundamental design decisions have been made, the design task generally follows the same course in step 2. The evaluator:

- Lists the questions which were raised by stakeholders and classifies them as requiring an Implementation, Progress or Summative Evaluation.
- Identifies procedures which might be used to answer these questions. Some of these procedures probably can be used to answer several questions; clearly, these will have priority.
- Looks at possible alternative methods, taking into account strength of the findings yielded by each approach (quality) as well as practical considerations especially time and cost constraints, staff availability, access to participants, etc.

An important consideration at this point is minimizing interference with project functioning.

An important consideration at this point is minimizing interference with project functioning: making as few demands as possible on project personnel and participants, and avoiding procedures which may be perceived as threatening or critical.

All in all, the evaluator will need to use a great deal of judgment in making choices and adjusting designs, and will seldom be in a position to fully implement text book recommendations. Some of the examples detailed in Chapter Six illustrate this point.

When and How to Sample

It is sometimes assumed that an evaluation must include all of the persons who participate in a project. Thus in teacher enhancement programs, all teachers need to be surveyed or observed; in studies of instructional practices, all students need to be tested; and in studies of reform, all legislators need to be interviewed. This is not the case.

Sampling may be considered or necessary for qualitative and quantitative studies. For example, if a project is carried out in a large number of sites, the evaluator may decide to carry out a qualitative study in only one or a few or them. When planning a survey of project participants, the investigator may decide to sample the participant population, if it is large. Of course, if the project involves few participants, sampling is unnecessary and inappropriate.

When planning allocation of resources, evaluators should give priority to procedures which will reduce sample bias and response bias, rather than to the selection of larger samples.

For qualitative studies, purposeful sampling is often most appropriate. Purposeful sampling means that the evaluator will seek out the case or cases which are most likely to provide maximum information, rather than a "typical" or "representative" case. The goal of the qualitative evaluation is to obtain rich, in-depth information, rather than information from which generalizations about the entire project can be derived. For the latter goal a quantitative evaluation is needed.

For quantitative studies, some form of random sampling is the appropriate method. The easiest way of drawing random samples is to use a list of participants (or teachers, or classrooms, or sites), and select every 2nd or 5th or 10th name, depending on the size of the population and the desired sample size. A stratified sample may be drawn to insure sufficient numbers of rare units (for example, minority members, or schools serving low-income students).

The most common misconception about sampling is that large samples are the best way of obtaining accurate findings. While it is true that larger samples will reduce **sampling error** (the probability that if another sample of the same size were drawn, different results might be obtained), sampling error is the smallest of the three components of error which affect the soundness of sample designs. Two other errors—**sample bias** (primarily due to loss of sample units) and **response bias** (responses or observations which do not reflect “true” behavior, characteristics or atti-

tudes)—are much more likely to jeopardize validity of findings. (Sudman, 1976). When planning allocation of resources, evaluators should give priority to procedures which will reduce sample bias and response bias, rather than to the selection of larger samples.

Let's talk a little more about sample and response bias. Sample bias occurs most often because of non-response (selected respondents or units are not available or refuse to participate, or some answers and observations are incomplete). Response bias occurs because questions are misunderstood or poorly formulated, or because respondents deliberately equivocate (for example to protect the project being evaluated). In observations, the observer may misinterpret or miss what is happening. Exhibit 4 describes each type of bias and suggests some simple ways of minimizing them.

Exhibit 4

Three Types of Errors and Their Remedies

Type	Cause	Remedies
Sampling Error	Using a sample, not the entire population to be studied.	Larger samples—these reduce but do not eliminate sampling error.
Sample Bias	Some of those selected to participate did not do so or provided incomplete information.	Repeated attempts to reach non-respondents. Prompt and careful editing of completed instruments to obtain missing data; comparison of characteristics of non-respondents with those of respondents to describe any suspected differences that may exist.
Response Bias	Responses do not reflect "true" opinions or behaviors because questions were misunderstood or respondents chose not to tell the truth.	Careful pretesting of instruments to revise mis-understood, leading, or threatening questions. No remedy exists for deliberate equivocation in self-administered interviews, but it can be spotted by careful editing. In personal interviews, this bias can be reduced by a skilled interviewer.

Determining an adequate sample size sounds threatening, but is not as difficult as it might seem to be at first. Statisticians have computed recommended sample sizes for various populations. (See Fitz-Gibbon and Morris, 1987.) For practical purposes, however, in project evaluations, sample size is primarily determined by available resources, by the planned analyses, and by the need for credibility.

In making sampling decisions, the overriding consideration is that the actual selection must be done by random methods, which usually means selecting every nth case from listings of units (students, instructors, classrooms). Sudman (1976) emphasizes that there are many scientifically sound sampling methods which can be tailored to all budgets:

"In far too many cases, researchers are aware that powerful sampling methods are available, but believe they cannot use them because these methods are too difficult and expensive. Instead incredibly sloppy ad hoc procedures are invented, often with disastrous results."

Deciding How Many Times to Measure

For all types of evaluations (Implementation, Progress, and Summative) the evaluator must decide the frequency of data collection and the method to be used if multiple observations are needed.

For many purposes, it will be sufficient to collect data at one point in time; for others one time data collection may not be adequate. Implementation Evaluations may utilize either multiple or one-time data collections depending on the length of the project and any problems that may be uncovered along the way. For Summative Evaluations, a one-time data collection may be adequate to answer some evaluation questions: How many students enrolled in the project? How many were persisters versus dropouts? What were the most popular project activities? Usually, such data can be obtained from records. But impact measures are almost always measures of change. Has the project resulted in higher test scores? Have teachers adopted different teaching styles? Have students become more interested in considering science-related careers? In each of these cases, at a minimum two observations are needed: baseline (at project initiation) and at a later point, when the project has been operational long

Impact measures are almost always measures of change.

enough for possible change to occur.

Quantitative studies using data collected from the same population at different points in time are called **longitudinal studies**. They often present a dilemma for the evaluator. Conventional wisdom suggests that the correct way to measure change is the "panel method," by which data are obtained from the same individuals (students, teachers, parents, etc.) at different points in time. While longitudinal designs which require interviewing the same students or observing the same teachers at several points in time are best, they are often difficult and expensive to carry out because students move, teachers are re-assigned, and testing programs are changed. Furthermore loss of respondents due to failure to locate or to obtain cooperation from some segment of the original sample is often a major problem. Depending on the nature of the evaluation, it may be possible to obtain good results with successive cross-sectional designs, which means drawing new samples for successive data collections from the treatment population. (See Love, 1991 for a fuller discussion of logistics problems in longitudinal designs.)

There is no hard and fast rule for deciding when changes should or should not be made; in the end technical concerns must be balanced with common sense.

For example, to evaluate the impact of a program of field trips to museums and science centers for 300 high school students, baseline interviews can be conducted with a random sample of 100 students before the project start. Interviewing another random sample of 100 students after the project has been operational for one year is an acceptable technique for measuring project effectiveness, provided that at both times samples were randomly selected to adequately represent the entire group of students involved in the project. In other cases, this may be impossible.

Designs that involve repeated data collection usually require that the data be collected using identical survey instruments at all times. Changing question wording or formats or observation schedules between time 1 and time 2 impairs the validity of the time comparison. At times, evaluators find after the first round of data collection that their instruments would be improved by making some changes, but they do so at the risk of not being able to use altered items for measuring change. Depending on the particular circumstances, it may be difficult to sort out whether a changed response is a treatment effect or the effect of the modified wording. There is no hard and fast rule

for deciding when changes should or should not be made; in the end technical concerns must be balanced with common sense.

How Do You Choose a Specific Data Collection Technique?

In Chapter Two we provided an overview of ways in which evaluators can go about collecting data. As shown in that chapter, there are many different ways to go about answering the same questions. However, the great majority of evaluation designs for projects supported by NSF/EHR rely at least in part on quantitative methods using one or several of the following techniques:

- Surveys based on self-administered questionnaires or interviewer administered instruments
- Focus groups
- Results from tests given to students
- Observations (most often carried out in classrooms)
- Review of records and data bases (not created primarily for the evaluation needs of the project).

The discussion in this section focuses on these techniques. Evaluators who are interested in using techniques not discussed here (for example designs using unobtrusive measures or videotaped observations) will find relevant information in some of the reference books cited in the bibliography.

Surveys

Surveys are a popular tool for project evaluation. They are especially useful for obtaining information about opinions and attitudes of participants or other relevant informants, but they are also useful for the collection of descriptive data, for example personal and background characteristics (race, gender, socio-economic status) of participants. Survey findings usually lend themselves to quantitative analysis; as in opinion polls, the results can be expressed in easily understood percentages or means. As compared to some other data collection methods, (for example in-depth interviews or observations) surveys usually provide wider ranging

but less detailed data and some data may be biased if respondents are not truthful. However, much has been learned in recent years about improving survey quality and coverage and compared to more intensive methods, surveys are relatively inexpensive and easier to analyze using statistical software.

The cheapest surveys are self-administered: a questionnaire is distributed (in person or by mail) to eligible respondents. Relatively short and simple questionnaires lend themselves best to this treatment. The main problem is usually non-response: persons not present when the questionnaire is distributed are often excluded, and mail questionnaires will yield relatively low response rates, unless a great deal of careful preparation and follow-up work is done.

When answers to more numerous and more complex questions are needed, it is best to avoid self-administered questionnaires and to employ interviewers to ask questions either in a face to face situation or over the telephone. Survey researchers often differentiate between questionnaires, where a series of precisely worded questions are asked, and interviews which are usually more open-ended, based on an interview guide or protocol and yield richer and often more interesting data. The trade-off is that interviews take longer, are best done face-to-face, and yield data which are often difficult to analyze. A good compromise is a structured questionnaire which provides some opportunity for open-ended answers and comments.

The choice between telephone and personal interviews depends largely on the nature of the projects being evaluated and the characteristics of respondents. For example, as a rule children should be interviewed in person, as should be respondents who do not speak English, even if the interview is conducted by a bilingual interviewer.

Creating a good questionnaire or interview instrument requires considerable knowledge and skill. Question wording and sequencing are very important in obtaining valid results, as shown by many studies. For a fuller discussion, see Fowler (1993, ch. 6) and Love (1991, ch. 2).

Focus groups

Focus groups have become an increasingly popular information gathering technique. Prior to designing survey instruments, a number of persons from the

population to be surveyed are brought together to discuss, with the help of a leader, the topics which are relevant to the evaluation and should be included in developing questionnaires. Terminology, comprehension, and recall problems will surface, which should be taken into account when questionnaires or interview guides are constructed. This is the main role for focus groups in Summative Evaluations. However, there may be a more substantive role for focus groups in Progress Evaluations, which are more descriptive in nature and often do not rely on statistical analyses. (See Stewart and Shamdasani, 1990 for a full discussion of focus groups.)

The usefulness of focus groups depends heavily on the skills of the moderator, the method of participant selection and last, but not least, the understanding of evaluators that focus groups are essentially an exercise in group dynamics. Their popularity is high because they are a relatively inexpensive and quick information tool, but while they are very helpful in the survey design phase, they are no substitute for systematic evaluation procedures.

Test Scores

Many evaluators and program managers feel that if a project has been funded to improve the academic skills of students so that they are prepared to enter scientific and technical occupations, improvements in test scores are the best indicator of a project's success. Test scores are often considered "hard" and therefore presumably objective data, more valid than other types of measurements such as opinion and attitude data, or grades obtained by students. But these views are not unanimous, since some students and adults are poor test-takers, and because some tests are poorly designed and measure the skills of some groups, especially White males, better than those of women and minorities.

Until recently, most achievement tests were either **norm-referenced** (measuring how a given student performed compared to a previously tested population) or **criterion-referenced** (measuring if a student had mastered specific instructional objectives and thus acquired specific knowledge and skills). Most school systems use these types of tests, and it has frequently been possible for evaluators to use data routinely collected in the schools as the basis for their summative studies.

Because of the many criticisms which have been directed at tests currently in use, there is now a great deal of interest in making radical changes. Experiments with **performance assessment** are under way in many states and communities. Performance tests are designed to measure problem solving behaviors, rather than factual knowledge. Instead of answering true/false or multiple choice formats, students are asked to solve more complex problems, and to explain how they go about arriving at answers and solving these problems. Testing may involve group as well as individual activities, and may appear more like a project than a traditional "test." While many educators and researchers are enthusiastic about these new assessments, it is not likely that valid and inexpensive versions of these tests will be ready for widespread use in the near future.

A good source of information about test vendors and for the use of currently available tests in evaluation is Morris, Fitz-Gibbon and Lindheim (1987). An extensive discussion of performance-based assessment by Linn, Baker, and Dunbar can be found in *Educational Researcher* (Nov. 1991).

Whatever type of test used, there are two critical questions that must be considered before selecting a test and using its results:

- Is there a match between what the test measures and what the project intends to teach? If a science curriculum is oriented toward teaching process skills, does the test measure these skills or more concrete scientific facts?

- Has the program been in place long enough for there to be an impact on test scores? With most projects, there is a start-up period during which the intervention is not fully in place. Looking for test score improvements before a project is fully established can lead to erroneous conclusions.

A final note on testing and test selection. Evaluators may be tempted to develop their own test instruments rather than relying on ones that exist. While this may at times be the best choice, it is not an option to be undertaken lightly. Test development is more than writing down a series of questions, and there are some strict standards formulated by the American Psychological Association that need to be met in developing instruments that will be credible in an evaluation. If at

all possible, use of a reliable and validated, established test is best.

Observations

Surveys and tests can provide good measurements of the opinions, attitudes, skills, and knowledge of individuals; surveys can also provide information about **individual behavior** (how often do you go to your local library? what did you eat for breakfast this morning?), but behavioral information is often inaccurate due to faulty recall or the desire to present oneself in a favorable light. When it comes to measuring **group behavior** (did most children ask questions during the science lesson? did they work cooperatively? at which museum exhibits did the students spend most of their time?) systematic observations are the best method for obtaining good data.

Evaluation experts distinguish between three observation procedures: (1) systematic observations, (2) anecdotal records (semi-structured), and (3) observation by experts (unstructured). For NSF/EHR project evaluations, the first and second are most frequently used, with the second to be used as a planning step for the development of systematic observation instruments.

Procedure one yields quantitative information, which can be analyzed by statistical methods. To carry out such quantifiable observations, subject-specific instruments will need to be created by the evaluator to fit the specific evaluation. A good source of information about observation procedures, including suggestions for instrument development, can be found in Henerson, Morris and Fitz-Gibbon (1987, ch. 9).

The main disadvantage of the observation technique is that behaviors may change when observed. This may be especially true when it comes to teachers and others who feel that the observation is in effect carried out for the purpose of evaluating their performance, rather than the project's general functioning. But behavior changes for other reasons as well, as noted a long time ago when the "Hawthorne effect" was first reported. Techniques have been developed to deal with the biasing effect of the presence of observers: for example, studies have used participant observers, but such techniques can only be used if the study does not call for systematically recording observations as events

occur. Another possible drawback is that perhaps more than any other data collection method, the observation method is heavily dependent on the training and skills of data collectors. This topic is more fully discussed later in this chapter.

Review of Records and Data Bases

Most agencies and funded projects maintain systematic records of some kind about the population they serve and the services they provide, but the extent of available information and their accessibility differ widely. The existence of a comprehensive Management Information System or data base is of enormous help in answering certain evaluation questions which in their absence may require special surveys. For example, simply by looking at personal characteristics of project participants, such as sex, ethnicity, family status etc. evaluators can judge the extent to which the project recruited the target populations described in the project application. As mentioned earlier, detailed project records will greatly facilitate the drawing of samples for various evaluation procedures. Project records can also identify problem situations or events (for example exceptionally high drop-out rates at one site of a multi-site project, or high staff turnover) which might point the evaluator in new directions.

Existing data bases which were originally set up for other purposes can also play a very important role in conducting evaluations. For example, if the project involves students enrolled in public or private institutions which keep comprehensive and/or computerized files, this would greatly facilitate the selection of "matched" control or comparison groups for complex outcome designs. However, gaining access to such information may at times be difficult because of rules designed to protect data confidentiality.

Exhibit 5 summarizes the advantages and drawbacks of the various data collection procedures.

What are Some Major Concerns When Collecting Data?

It is not possible to discuss in one brief chapter the nitty-gritty of all data collection procedures. The reader will want to consult one or more of the texts recommended in the bibliography before attacking any one specific task. Before concluding this chapter, we want to address two issues, however, which affect all data

Exhibit 5

Advantages and Drawbacks of Various Data Collection Procedures		
Procedure	Advantages	Disadvantages
Self-administered questionnaire	Inexpensive. Can be quickly administered if distributed to group. Well suited for simple and short questionnaires.	No control for misunderstood questions, missing data, or untruthful responses. Not suited for exploration of complex issues.
Interviewer administered questionnaires (by telephone)	Relatively inexpensive. Avoids sending staff to unsafe neighborhoods or difficulties gaining access to buildings with security arrangements. Best suited for relatively short and non-sensitive topics.	Proportion of respondents without a private telephone may be high in some populations. As a rule not suitable for children, older people, and non-English speaking persons. Not suitable for lengthy questionnaires and sensitive topics. Respondents may lack privacy.
Interviewer administered questionnaires (in person)	Interviewer controls situation, can probe irrelevant or evasive answers; with good rapport, may obtain useful open-ended comments.	Expensive. May present logistics problems (time, place, privacy, access, safety). Often requires lengthy data collection period unless project employs large interviewer staff.
Open-ended interviews (in person)	Usually yields richest data, details, new insights. Best if in-depth information is wanted.	Same as above (interviewer administered questionnaires); also often difficult to analyze.
Focus groups	Useful to gather ideas, different viewpoints, new insights, improving question design.	Not suited for generalizations about population being studied.
Tests	Provide "hard" data which administrators and funding agencies often prefer; relatively easy to administer; good instruments may be available from vendors.	Available instruments may be unsuitable for treatment population; developing and validating new, project-specific tests may be expensive and time consuming. Objections may be raised because of test unfairness or bias.
Observations	If well executed, best for obtaining data about behavior of individuals and groups.	Usually expensive. Needs well qualified staff. Observation may affect behavior being studied.

collections and deserve special mention here: the selection, training, and supervision of data collectors, and pretesting of evaluation instruments.

Selection, Training and Supervision of Data Collection

Selection

All too often, project administrators, and even evaluators, believe that anybody can be a data collector and typically base the selection on convenience factors: an available research assistant, an instructor or clerk willing to work overtime, college students available for part-time or sporadic work assignments. All of these may be suitable candidates, but it is unlikely that they will be right for all data collection tasks.

Most data collection assignments fall into one of three categories:

- Clerical tasks (abstracting records, compiling data from existing lists or data bases, keeping track of self-administered surveys)
- Personal interviewing (face-to-face or by telephone) and test administration
- Observing and recording observations.

There are some common requirements for the successful completion of all of these tasks: a good understanding of the project, ability and discipline to follow instructions consistently and to give punctilious and detailed attention to all aspects of the data collection. Equally important is lack of bias, and lack of vested interest in the outcome of the evaluation. For this reason, as previously mentioned (Chapter Two) it is usually unwise to use volunteers or regular project staff as data collectors.

Interviewers need additional qualities: a pleasant voice and tactful personal manner and the ability to establish rapport with respondents. For some data collections, it may be advisable to attempt a match between interviewer and respondent (for example with respect to ethnicity, or age.) The need for fluency in a language other than English (usually Spanish) may also be needed; in this case it is important that the interviewer be bi-lingual, with U.S. work experience, so that instructions and expected performance stan-

dards are well understood.

Observers need to be highly skilled and competent professionals. Although they too will need to follow instructions and complete structured schedules, it is often important that they alert the evaluator to unanticipated developments. Depending on the nature of the evaluation, their role in generating information may be crucial: often they are the eyes and the ears of the evaluator. They should also be familiar with the setting in which the observations take place, so that they know what to look for. For example teachers (or former teachers or aides) can make good classroom observers, although they should not be used in schools with which they are or were affiliated.

Training

In all cases, sufficient time must be allocated to training. Training sessions should include performing the actual task (extracting information from a data base, conducting an interview, performing an observation). Training techniques might include role-playing (for interviews) or comparing recorded observations of the same event by different observers. When the project enters a new phase (for example when a second round of data collection starts) it is usually advisable to schedule another training session, and to check inter-rater reliability again.

If funds and technical resources are available, other techniques (for example videotaping of personal interviews or recording of telephone interviews) can also be used for training and quality control after permission has been obtained from participants.

Supervision

Only constant supervision will ensure quality control of the data collection. The biggest problem is not cheating by interviewers or observers (although this can never be ruled out), but gradual burnout: more transcription errors, more missing data, fewer probes or follow-ups, fewer open-ended comments on observation schedules.

The project evaluator should not wait to review completed work until the end of the data collection, but should do so at least once a week. See Fowler (1991) and Henerson, Morris and Fitz-Gibbon (1987) for further suggestions on interviewer and observer re-

cruitment and training.

Pretest of Instruments

When the evaluator is satisfied with the instruments designed for the evaluation, and before starting any data collection in the field, all instruments should be pre-tested to see if they work well under field conditions. The pre-test also reveals if questions are understood by respondents and if they capture the information sought by the evaluator. Pre-testing is a step that many evaluators "skip" because of time pressures. However, as has been shown many times, they may do so at their own peril. The time taken up front to pre-test instruments can result in enormous savings in time (and misery) later on.

Pre-testing is a step that many evaluators "skip" because of time pressures. However, as has been shown many times, they may do so at their own peril.

The usual procedure consists of using instruments with a small number of cases (for example abstracting data from 10 records, asking 10-20 project participants to fill out questionnaires, conducting interviews with 5 to 10 subjects, or completing half a dozen class room observations). Some of the shortcomings of the instruments will be obvious as the completed forms are reviewed, but most important is a debriefing session with data collectors and in some instances with the respondents themselves, so that they can recommend to the evaluator possible modifications of procedures and instruments. It is especially important to pre-test self-administered instruments, where the respondent cannot ask an interviewer for help in understanding questions. Such pre-tests are best done by bringing together a group of respondents, asking them first to complete the questionnaire, and then leading a discussion about clarity of instructions, and understanding the questions and expected answers.

Data Analysis: Qualitative Data

Analyzing the plethora of data yielded by comprehensive qualitative evaluations is a difficult task, and there are many instances of frequent failure to fully analyze the results of long and costly data collections. While lengthy descriptive case studies are extremely useful in furthering the understanding of social phenomena and the implementation and functioning of innovative projects, they are ill-suited to outcome evaluation studies for program managers and funding agencies. However, more recently, methods have been devised to classify qualitative findings through the use of a special software program (Ethnograph) and di-

verse thematic codes. This approach may enable investigators to analyze qualitative data quantitatively without sacrificing the richness and character of qualitative analysis. Content analysis which can be used for the analysis of unstructured verbal data, is another available technique for dealing quantitatively with qualitative data. Other approaches, including some which also seek to quantify the descriptive elements of case studies, and others which address issues of validation and verification also suggest that the gap between qualitative and quantitative analyses is narrowing. Specific techniques for the analysis of qualitative data can be found in some of the texts referenced at the end of this Chapter.

Data Analysis: Quantitative Data

In Chapter Two, we outlined the major steps required for the analysis of quantitative data:

- Check the raw data and prepare data for analysis
- Conduct initial analysis based on evaluation plan
- Conduct additional analyses based on initial results
- Integrate and synthesize findings.

In this chapter, we provide some additional advice on carrying out these steps.

Check the Raw Data and Prepare Data for Analysis

In almost all instances, the evaluator will conduct the data analysis with the help of a computer. Even if the number of cases is small, the volume of data collected and the need for accuracy, together with the availability of PC's and user-friendly software, make it unlikely that evaluators will do without computer assistance.

The process of preparing data for computer analysis involves **data checking**, **data reduction**, and **data cleaning.**

Data checking can be done as a first step by visual inspection of the raw data; this check may turn up responses which are out-of-line, unlikely, inconsistent or suggest that a respondent answered questions mechanically (for example chose always the third response category in a self-administered question-

naire).

Data reduction consists of the following steps:

- Deciding on a file format. (This is usually determined by the software to be used.)
- Designing codes (the categories used to classify the data so that they can be processed by machine) and coding the data. If instruments are "pre-coded," for example if respondents were asked to select an item from a checklist, coding is not necessary. It is needed for "open-ended" answers and comments by respondents and observers.
- Data entry (keying the data onto tapes or disks so that the computer can read them).

Solid data preparation procedures help avoid "GIGO"- garbage in, garbage out.

Many quality control procedures for coding open-ended data and data entry have been devised. They include careful training of coders, frequent checking of their work, and verification of data entry by a second clerk.

Data cleaning consists of a final check on the data file for accuracy, completeness and consistency. At this point, coding and keying errors will be detected. (For a fuller discussion of data preparation procedures, see Fowler, 1991).

If these data preparation procedures have been carefully carried out, chances are good that the data sets will be error-free from a technical standpoint and that the evaluator will have avoided the "GIGO" (garbage in, garbage out) problem which is far from uncommon in analyses based on computer output.

Conduct Initial Analysis Based on the Evaluation Plan

In fact, much can be learned from fairly uncomplicated techniques easily mastered by persons without a strong background in mathematics or statistics.

The evaluator is now ready to start generating information which will answer the evaluation questions. To do so, it is usually necessary to deal with statistical concepts and measurements, a prospect which some evaluators or principal investigators may find terrifying. In fact, much can be learned from fairly uncomplicated techniques easily mastered by persons without a strong background in mathematics or statistics. Many evaluation questions can be answered through the use of descriptive statistical measures, such as frequency distributions (how many cases fall into a

given category), and measures of central tendency (such as the mean or median which refer to statistical measures which seek to locate the "average" or the center of a distribution).

For frequency distributions, the question is most often a matter of presenting such data in the most useful form for project managers and stakeholders. Often the evaluator will look at detailed distributions and then decide on a summary presentation, using tables or graphics. An example is the best way of illustrating these various issues.

Let us assume that a project had recruited 200 high school students to meet with a mentor once a week over a one year period. One of the evaluation questions was: "How long did the original participants remain in the program?" Let us also assume that the data were entered in weeks. If we ask the computer to give us a frequency distribution, we get a long list (if every week at least one participant dropped out, we may end up with 52 entries for 200 cases). Eyeballing this unwieldy table, the evaluator noticed several interesting features: only 50 participants (1/4th of the total) stayed for the entire length of the program; a few people never showed up or stayed only for 1 session. To answer the evaluation question in a meaningful way, the evaluator decided to ask the computer to group data into a shorter table, as follows:

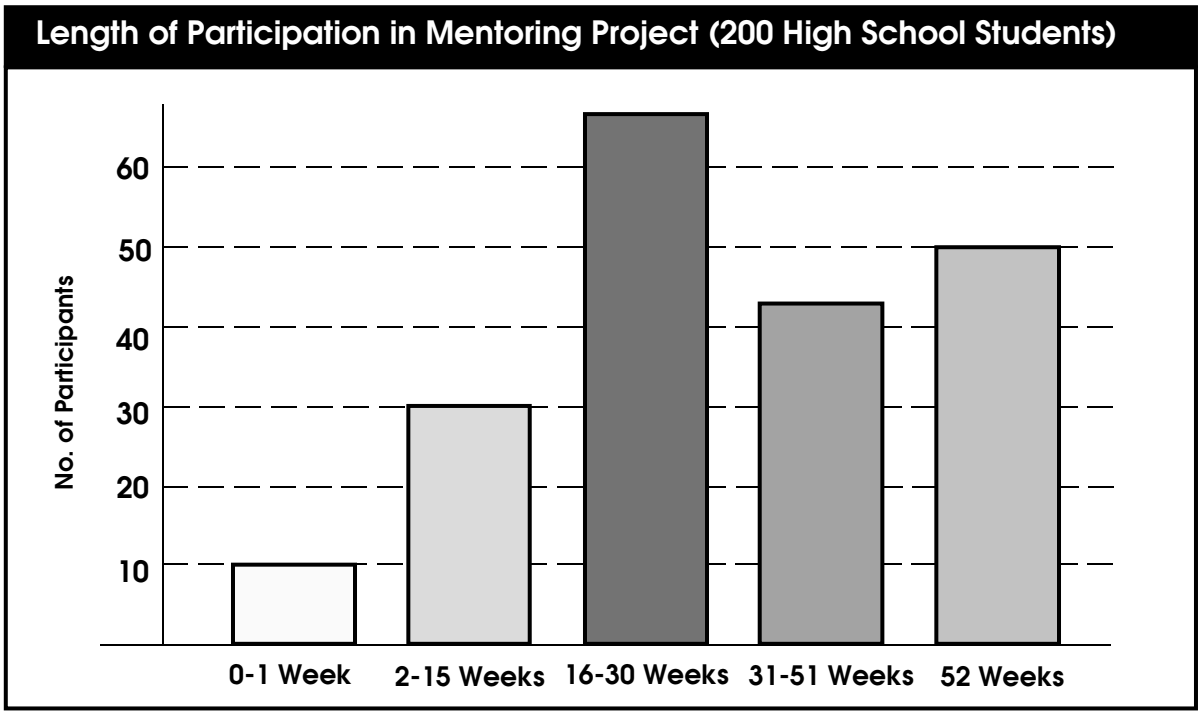
Length of Participation

Time	No. of Participants
1 week or less	10
2-15 weeks	30
16-30 weeks	66
31-51 weeks	44
52 weeks	50

A bar chart might be another way of presenting these data as shown in Exhibit 6.

Let us now assume that the evaluator would like a single figure which would provide some indication of the length of time during which participants remained in the project. There are three measures of central tendency which provide this answer, the mean (or

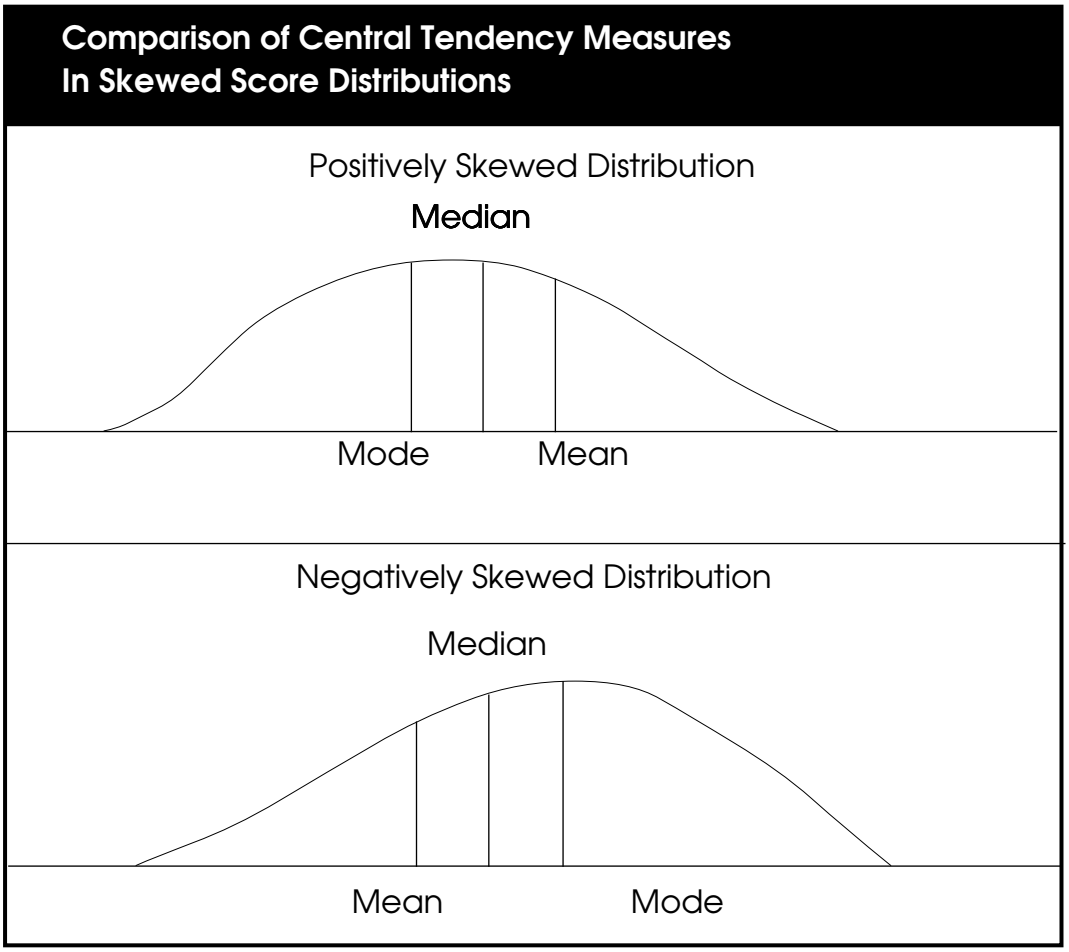
Exhibit 6



arithmetic average), the median (the point at which half the cases fall below and half above), and the mode, which is the category with the largest number of cases. Each of these require that the data meet specific conditions and each has advantages and drawbacks. (See glossary for details.)

In the above example, the only way of computing the mean, median, and mode would be from the raw data, prior to grouping the data as shown in Exhibit 7. However, to simplify the discussion we will just deal with the mean and median (usually the most meaningful measures for evaluation purposes), which can be computed from grouped data. The mean would be slightly above 30 weeks, the median would be slightly above 28 weeks. The mean is higher because of the impact of the last two categories (31-51 weeks and 52 weeks). Both measures are "correct," but they tell slightly different things about the length of time participants remained in the project; the average was 30 weeks, which may be a useful figure for estimating future project costs; half of all participants stayed for 28 weeks or less, which may be a useful figure for deciding how to time retention efforts. Exhibit 7 illustrates differences in the relative position of the median, mean, and mode depending on the nature of

Exhibit 7



the data, such as a positively skewed distribution of test scores (more test scores at the lower end of the distribution) and for a negatively skewed distribution (more scores at the higher end).

In many evaluation studies, the median is the preferred measure of central tendency because for most analyses, it describes the distribution of the data better than the mode or the mean. For a useful discussion of these issues, see Jaeger (1990).

Conduct Additional Analyses Based on the Initial Results

The initial analysis may give the evaluator a good feel for project operations, levels of participation, project activities, and the opinions and attitudes of partici-

pants, staff, and others involved in the project, but it often raises new questions. These may be answered by additional analyses which examine the findings in greater detail. For example, as discussed in some of the earlier examples, it may be of interest to compare more and less experienced teachers' assessment of the effectiveness of new teaching materials, or to compare the opinions of men and women who participated in a mentoring program. Or it might be useful to compare the opinions of women who had female mentors with those of women who had male mentors. These more detailed analyses are often based on **cross-tabulations,** which, unlike frequency distributions, deal with more than one variable. If, in the earlier example about length of participation in mentoring programs, the evaluator wants to compare men and women, the cross-tabulation would look as follows:

One rule of thumb is that a minimum of 20 cases are needed in each subgroup for analysis and for the use of statistical tests to judge the extent to which observed differences are "real" or due to sampling error.

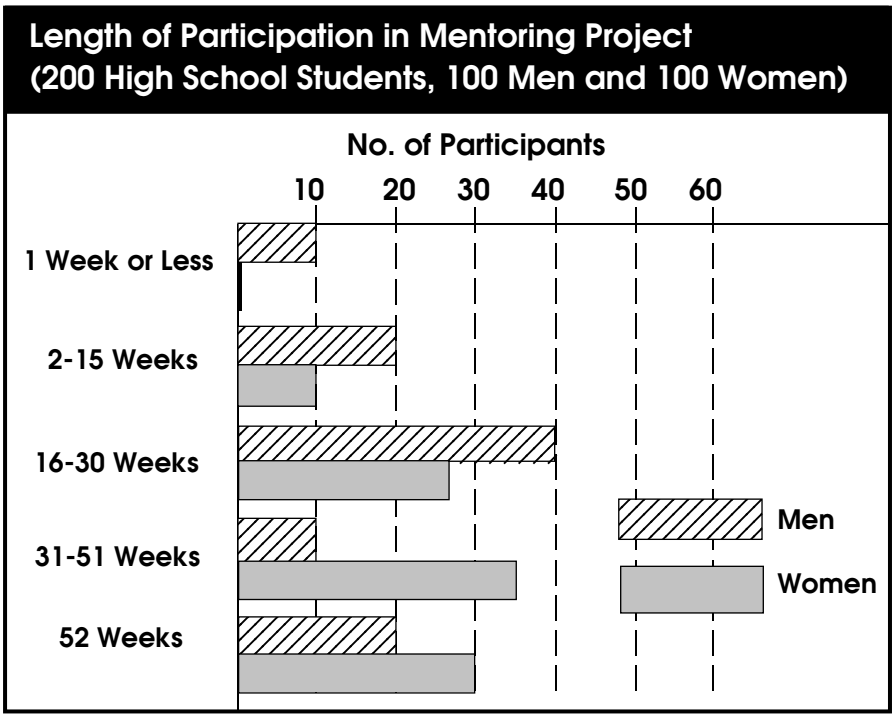
Length of Participation by Sex

	All Students	Men	Women
1 week or less	10	10	0
2-15 weeks	30	20	10
16-30 weeks	66	40	26
31-51 weeks	44	10	34
52 weeks	50	20	30

Exhibit 8, a bar graph, is a better way of showing the same data. Because the table and graph show that on the whole women dropped out later than men, but that most of them also did not complete the entire program, the evaluator may want to re-group the data, for example break down the 31-51 group further to see if most women stayed close to the end of the program.

Cross-tabulations are a convenient technique for examining several variables simultaneously; however, they are often inappropriate because sub-groups become too small. One rule of thumb is that a minimum of 20 cases are needed in each subgroup for analysis and for the use of statistical tests to judge the extent to which observed differences are "real" or due to sampling error. In the above example, it might have been of interest to look further at men and women in different ethnic groups (African American men, African American women, White men and White women) but among the 200 participants there might not have been a sufficient number of African American men or White women to carry out the analysis.

Exhibit 8



There are other techniques for examining differences between groups and testing the findings to see if the observed differences are likely to be "true" ones. To use any one of them, the data must meet specific conditions. Correlation, t-tests, chi-square, and variance analysis are among the most frequently used and have been incorporated in many standard statistical packages. More complex procedures, designed to examine a large number of variables and measure their respective importance, such as factor analysis, regression analysis, and analysis of co-variance are powerful statistical tools, but their use requires a higher level of statistical knowledge. There are special techniques for the analysis of longitudinal (panel) data. Many excellent sources are available for deciding about the appropriateness and usefulness of the various statistical methods (Jaeger, 1990; Fitz-Gibbon and Morris, 1987).

Exploring the data by various statistical procedures in order to detect new relationships and unanticipated findings is perhaps the most exciting and gratifying evaluation task.

Exploring the data by various statistical procedures in order to detect new relationships and unanticipated findings is perhaps the most exciting and gratifying evaluation task. It is often rewarding and useful to keep exploring new leads, but the evaluator must not lose track of time and money constraints and needs to recognize when the point of diminishing returns has been reached.

By following the suggestions made so far in this chapter, the evaluator will be able to answer many questions about the project asked by stakeholders concerned about implementation, progress, and some outcomes. But the question most often asked by funding agencies, planners and policy makers who might want to replicate a project in a new setting is the question: Did the program achieve its objectives? Did it work? What feature(s) of the project were responsible for its success or failure? Outcome evaluation is the evaluator's most difficult task. It is especially difficult for an evaluator who is familiar with the conceptual and statistical pitfalls associated with program evaluation. To quote what is considered by many the classic text in the field of evaluation (Rossi and Freeman, 1993):

"The choice (of designs) always involves trade-offs, there is no single, always-best design that can be used as the 'gold standard'."

Why is outcome evaluation or impact assessment so difficult? The answer is simply that educational projects do not operate in a laboratory setting, where "pure" experiments can yield reliable findings pointing to cause and effect. If two mice from the same litter are fed different vitamins, and one grows faster than the other, it is easy to conclude that vitamin x affected growth more than vitamin y. Some projects will try to measure impact of educational innovations by using this scientific model: observing and measuring outcomes for a treatment group and a matched comparison group. While such designs are best in theory, they are by no means fool-proof: the literature abounds in stories about "contaminated" control groups. For example, there are many stories about teachers whose students were to be controls for an innovative program, and who made special efforts with their students so that their traditional teaching style would yield exceptionally good outcomes. In other cases, students in a control group were subsequently enrolled in another experimental project. But even if the control group is not contaminated, there are innumerable questions about attributing favorable outcomes to a given project. The list of possible impediments is formidable. Most often cited is the fallacy of equating high correlation with causality. If attendance in the mentoring program correlated with higher test scores, was it because the program stimulated the students to study harder and helped them to understand scientific concepts better? Or was it because those who

chose to participate were more interested in science than their peers? Or was it because the school changed its academic curriculum requirements? Besides poor design and measurements, the list of factors which might lead to spurious outcome assessments includes invalid outcome measures as well as competing explanations, such as changes in the environment in which the project operated and Hawthorne effects. On the basis of his many years of experience in evaluation work, Rossi and Freeman (1993) formulated 'The Iron Law of Evaluation Studies':

"The better an evaluation study is technically, the less likely it is to show positive program effects."

There is no formula which can guarantee a flawless and definitive outcome assessment. Together with a command of analytic and statistical methods, the evaluator needs the ability to view the project in its larger context (the real world of real people) in order to make informed judgments about outcomes which can be attributed to project activities. And, at the risk of disappointing stakeholders and funding agencies, the evaluator must stick to his guns if he feels that available data do not enable him to give an unqualified or positive outcome assessment. This issue is further discussed in Chapter Four.

Integrate and Synthesize Findings

When the data analysis has been completed, the final task is to select and integrate tables, graphs and figures which constitute the salient findings and will provide the basis for the final report. Usually the evaluator must deal with several dilemmas:

- How much data must be presented to support a conclusion?
- Should data be included which are interesting or provocative, but do not answer the original evaluation questions?
- What to do about inconsistent or contradictory findings?

Here again, there are no hard and fast rules. Because usually the evaluator will have much more information than can be presented, judicious selection should guide the process. It is usually unnecessary to belabor

a point by showing all the data on which the conclusion is based: just show the strongest indicator. On the other hand, "interesting" data which do not answer one of the original evaluation questions should be shown if they will help stakeholders to understand or seek to address issues of which they may not have been aware. A narrow focus of the evaluation may fulfill contractual or formal obligations, but it deprives the evaluator of the opportunity to demonstrate substantive expertise and the stakeholders of the full benefit of the evaluator's work. Finally, inconsistent or contradictory findings should be carefully examined to make sure that they are not due to data collection or analytic errors. If this is not the case, they should be put on the table, as pointing to issues which may need further thought or examination.

REFERENCES

American Psychological Association, Educational Research Association, and National Council on Measurement in Education (1974). *Standards for Educational and Psychological Tests*. Washington, DC: American Psychological Association.

Fitz-Gibbon, C. T. & Morris, L. L. (1987). *How to Design a Program Evaluation.* Newbury Park, CA: Sage.

Fowler, F. J. (1993). S*urvey Research Methods.* Newbury Park, CA: Sage.

Guba, E. G. & Lincoln, Y. S. (1989). *Fourth Generation Evaluation*. Newbury Park, CA: Sage.

Henerson, M. E., Morris, L. L., & Fitz-Gibbon, C. T. (1987). *How to Measure Attitudes.* Newbury Park, CA: Sage.

Herman, J. L., Morris, L. L., & Fitz-Gibbon, C. T. (1987). *Evaluators Handbook.* Newbury Park, CA: Sage.

Jaeger, R. M. (1990). *Statistics—A Spectator Sport.* Newbury Park, CA: Sage.

Linn, R. L., Baker, E. L., & Dunbar, S. B. "Complex performance-based assessment: expectations and validation criteria." *Educational Researcher*, 20-8, 1991.

Love, A. J. (ed.) (1991). *Evaluation Methods Sourcebook.* Ottawa, Canada: Canadian Evaluation Society.

Morris, L. L., Fitz-Gibbon, C. T., & Lindheim, E. (1987). *How To Measure Performance and Use Tests.* Newbury Park, CA: Sage.

Rossi, P. H. & Freeman, H .E. (1993). *Evaluation—A Systematic Approach* (5th Edition). Newbury Park, CA: Sage.

Scriven, M. (1991). *Evaluation Thesaurus.* Newbury Park, CA: Sage.

Seidel, J. V., Kjolseth, R. & Clark, J. A. (1988). *The Ethnograph.* Littleton, CO: Qualis Research Associates.

Stewart, P. W. & Shamdasani, P. N. (1990). *Focus Groups.* Newbury Park, CA : Sage.

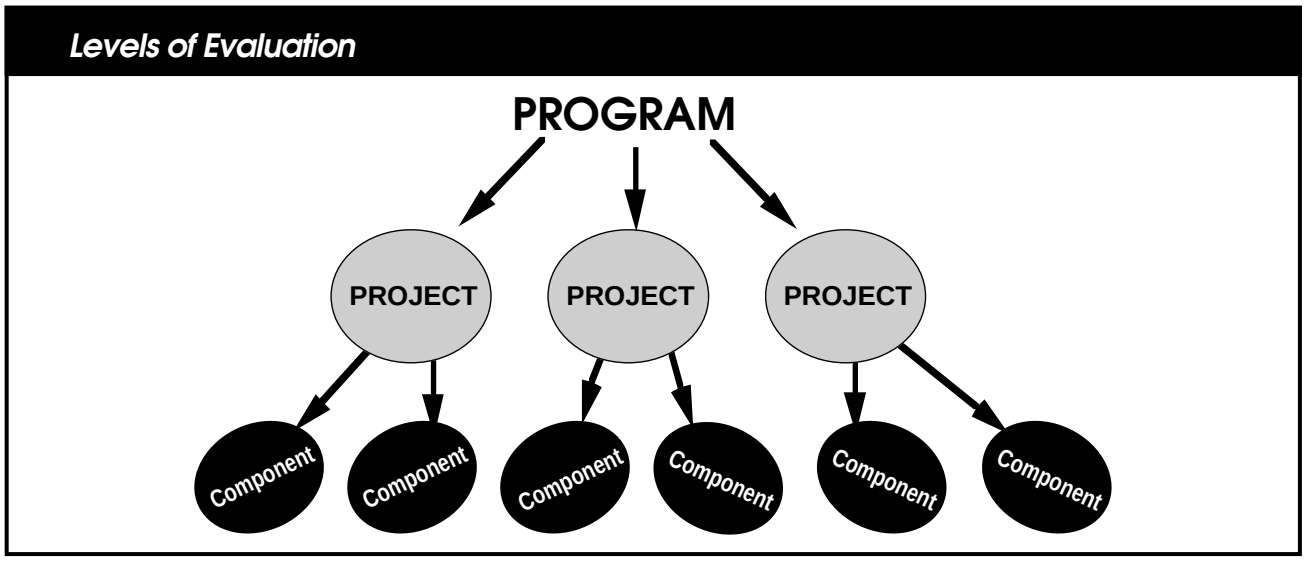
Sudman, S. (1976). *Applied Sampling.* New York: Academic Press.

Yin, R. (1989). *Case Study Research.* Newbury Park, CA: Sage.

What are the Different Kinds of Evaluations?

Within NSF, there are two basic levels of evaluation: Program Evaluation and Project Evaluation. Project Evaluation is sometimes further subdivided into specific project components as shown in Exhibit 1.

Exhibit 1



Let's start by defining terms and showing how they relate. First, let's define what we mean by a "program" and a "project." A **program** is a coordinated approach to exploring a specific area related to NSF's mission of strengthening science, mathematics, engineering, and technology. A **project** is a particular investigative or developmental activity funded by that program. NSF initiates a program on the assumption that a policy goal (for example, strengthening minority student development) can be attained by certain educational activities and strategies (for example, exposing students in inner-city schools to science presentations targeted at the interests and concerns of young African Americans). The Foundation then funds a series of discrete projects to explore the utility of these activities and strategies in specific situations. Thus, a program consists of a collection of projects that seek to meet a defined set of goals and objectives.

Now, let's turn to the terms "Program" and "Project Evaluation." A **Program Evaluation** determines the value of this collection of projects. It looks across projects, examining the utility of the activities and strategies employed, in light of the initial policy goal. It is carried to completion after the projects have become fully operational and adequate time has passed for expected outcomes to be achieved. Frequently, the initiation of a Program Evaluation is deferred until 3 to 5 years after program initiation. Other times, reporting may be deferred, while data collection is begun simultaneously with program onset. Under this latter alternative, the evaluation could draw upon information collected on an annual basis that is aggregated across projects and summarized at an appropriate check point. The program evaluator is, in this case, usually an experienced, external evaluator, selected by NSF.

In NSF, a Program Evaluation determines the value of a collection of projects. Project Evaluation, in contrast, focuses on the individual projects funded under the umbrella of the program.

Project Evaluation, in contrast, focuses on an individual project funded under the umbrella of the program. The evaluation provides information to improve the project as it develops and progresses. Information is collected to help determine whether it is proceeding as planned; whether it is meeting its stated program goals and project objectives according to the proposed timeline. Frequently these evaluation findings are also used to assess whether the particular project merits continued funding as it is currently operating, or if it needs modifications. Ideally in a Project Evaluation, evaluation design and data collection begin soon after the project is funded. Data collection occurs on a planned schedule, e.g., every 6 months or every year; and may lead to and support recommendations to continue, modify, and/or delete project activities and strategies. Frequently, although not universally, the Project Evaluator is a member of the project staff, is selected by, and reports to the Project Director.

Project Evaluations may also include examination of specific components. A component of a project may be a specific teacher training approach, a classroom practice, or a governance strategy. An evaluation of a component frequently looks to see the extent to which its goals have been met (these goals are a subset of the overall project goals), and to clarify the extent to which the component contributes to the success or failure of the overall project.

The information contained in this Handbook has been primarily prepared for the use of Project Evaluators

and Principal Investigators, although Program Evaluators may also find it useful. Our aim is to provide tools that will help those responsible for examination of individual projects gain the most from their evaluation efforts. Clearly, however, these activities will also benefit program studies and the work of the Foundation in general. The better the information about each of the NSF projects, the more we can all learn.

In the next section we describe three general types of evaluation studies: (1) Planning Evaluation, (2) Formative (Implementation and Progress) Evaluation, and (3) Summative Evaluation. While Summative Evaluation is frequently the notion that comes to mind when the term "evaluation" is used, each has its own contribution to make in understanding how well a project is doing. As each type of evaluation is discussed we present a brief definition of its purpose and some ideas of the kinds of questions that could be addressed.

What is a Planning Evaluation?

The purpose of a Planning Evaluation is to assess understanding of a project's goals, objectives, strategies, and timelines. "Planning Evaluation" is not as commonly carried out as the other prototypes. In fact most project proposals typically mention only "Formative" and "Summative" Evaluation, defining these as activities to be performed once a project has been designed, written up, and funded. The evaluator enters the scene after the project has been put in place.

The purpose of a Planning Evaluation is to assess understanding of projects' goals, objectives, strategies, and timelines.

A strong argument can be made for a different approach. Rossi and Freeman (1993) argue strongly for the involvement of evaluators in diagnosing and defining the condition that a given project is designed to address, in stating clearly and precisely the goals of the project, and in reviewing the proposed procedures for accuracy of information and soundness of methods.

The Planning Evaluation will provide everyone—Program Directors, Principal Investigators, Project Directors/Managers, participants, and the public—with an understanding of what the project is supposed to do and the timelines and strategies for doing it. The product of the Planning Evaluation is a rich, context-laden description of a project, including its major goals and objectives, activities, participants and other major stakeholders, resources, timelines, locale, and intended accomplishments. The Planning Evaluation can also serve the purpose of describing the status of

key outcome indicators prior to the project to serve as a baseline for measuring success.

To conduct a Planning Evaluation, the evaluator should be present when the project is in its developmental phase. The Planning Evaluation is typically designed to address the following questions:

- Why was the project developed? What is the problem or need it is attempting to address?
- Who are the stakeholders (those who have credibility, power, or other capital involved in the project)? Who are the people interested in the project who may not be involved?
- What do the stakeholders want to know? What questions are most important to which stakeholders? What questions are secondary in importance? Where do concerns coincide? Where are they in conflict?
- Who are the participants to be served?
- What are the activities and strategies that will address the problem or need which was identified? What is the intervention? How will participants benefit? What are the expected outcomes?
- Where will the program be located (educational level, geographical area)?
- How many months of the school year or calendar year will the program operate? When will the program begin and end?
- How much does it cost? What is the budget for the program? What human, material, and institutional resources are needed? How much is needed for evaluation? for dissemination?
- What are the measurable outcomes which the project wants to achieve? What is the expected impact of the project in the short run? the longer run?
- What arrangements have been made for data collection? What are the understandings regarding record keeping, responding to surveys, and participation in testing?

These questions can become a checklist to determine

if all relevant elements are included in the description of the project or program. These questions also provide the basis for the formative and summative evaluative inquiries about the project.

What is Formative Evaluation?

The purpose of Formative Evaluation is to assess ongoing project activities. Formative Evaluation begins at project start-up and continues throughout the life of the project. Its intent is to provide information to improve the project. It is done at several points in the developmental life of a project. According to evaluation theorist Bob Stake, Formative Evaluation, when contrasted with Summative Evaluation, is:

The purpose of Formative Evaluation is to assess ongoing project activities.

"When the cook tastes the soup, that's formative; when the guests taste the soup, that's summative."

For most NSF projects, Formative Evaluation consists of two segments: Implementation Evaluation and Progress Evaluation.

What is Implementation Evaluation?

The purpose of an Implementation Evaluation is to assess whether the project is being conducted as planned. It may occur once or several times during the life of the project. Recall the principle learned from the tale of the emperor who had no clothes and no one would tell him. The same principle applies to a new project or new program. Before you can evaluate the outcomes of a project, you must make sure the project is really operating, and if it is operating according to its plan or description. For example, in the description for Comprehensive Regional Centers for Minorities (CRCM), these Regional Centers must be comprehensive in their coverage of science, engineering and mathematics and focus on a span of educational levels—elementary through high school. An Implementation Evaluation of a CRCM project might begin by investigating whether or not the CRCM was indeed comprehensive in its coverage and whether its focus spanned elementary through senior high school. If these two essential conditions were satisfied, it could be concluded that the CRCM was initially implemented as intended and that evaluation of outcomes and impacts associated with the

implementation could proceed.

Implementation Evaluation collects information to determine if the program or project is being delivered as planned. A series of implementation questions is needed to guide the Implementation Evaluation. Examples of these questions are:

- Were the appropriate participants selected and involved in the planned activities?
- Do the activities and strategies match those described in the plan? If not, are the changes in activities justified and described?
- Were the appropriate staff members hired, trained, and are they working in accordance with the proposed plan? Were the appropriate materials and equipment obtained?
- Were activities conducted according to the proposed timeline? by appropriate personnel?
- Was a management plan developed and followed?

The purpose of an Implementation Evaluation is to assess whether the project is being conducted as planned.

Sometimes the terms "Implementation Evaluation" and "Monitoring Evaluation" are confused. They are not the same. While Implementation Evaluation is an early internal check by the project staff to see if all the essential elements of the project are in place and operating, monitoring is an external check and should follow the Implementation Evaluation. The monitor comes from the funding agency and is responsible for determining progress and compliance on a contract or grant for the project. The monitor investigates proper use of funds, observes progress, and provides information to the funding agency about the project. Although the two differ, Implementation Evaluation, if effective, can facilitate and ensure that there are no unwelcome surprises during monitoring.

What is Progress Evaluation?

The purpose of a Progress Evaluation is to assess progress in meeting the project's goals. Progress Evaluation is also formative. It involves collecting information to learn whether or not the benchmarks of participant progress were attained and to point out unexpected developments. Progress Evaluation collects information to determine what the impact of the activities and strategies is on the participants at

various stages of the intervention. By measuring interim outcomes, project staff eliminate the risk of waiting until participants have experienced the entire treatment to assess outcomes. If the data collected as part of the Progress Evaluation fail to show expected changes, this information can be used to "fine-tune" or terminate the project. The data collected as part of a Progress Evaluation can also contribute to, or form the basis for, a Summative Evaluation study conducted at some future date. In a Progress Evaluation, the following questions could be asked:

- Are the participants moving toward the anticipated goals of the project or program?
- Which of the activities and strategies are aiding the participants to move toward the goals?

The purpose of a Progress Evaluation is to assess progress in meeting the project's goals.

For example, one of the goals for the Alliances for Minority Participation (AMP) Program is to increase the size of the pool of underrepresented minority students eligible for Science, Engineering, and Mathematics (SEM) graduate study. One of the interim indicators which shows progress towards meeting the goal is the number (percent) of participants in the Summer Bridge program (a component of the AMP Program) who successfully complete calculus by their freshman year in college. Additional progress information could be scores from calculus classroom quizzes throughout the summer before the final exam and grades given for the course. Collecting this information on course completion, test scores, and grades, gives interested parties some idea of the rate and extent to which progress is being made toward the overarching goal of increasing the numbers of underrepresented minority students eligible for SEM graduate study. It gives some idea of the probability of achieving that final goal. If course completion or other indicators are not showing progress, significant project changes may be considered.

Another example of measuring progress can be drawn from Comprehensive Regional Centers for Minorities (CRCM). A goal is that, through workshops, teachers will learn to improve and enrich their teaching strategies when teaching classes such as high school chemistry. This interim goal is related to meeting the CRCM goal of retaining precollege students' interest in science. Progress findings could include teachers' ratings of their inservice training classes, and the independent

appraisals by outside observers of the quality of their performance when using new strategies in the classroom. In addition, the opinions and attitudes of the participants (students and teachers) could be collected to determine whether the impact of the activities and strategies is negative or positive.

In Progress Evaluation, quantitative and qualitative information about the participants is collected to determine if parts of the project need to be changed or deleted to improve the project. Progress Evaluation is useful throughout the life of a project, but is most vital during the early stages when activities are piloted and their individual effectiveness or articulation with other project components is unknown.

What is Summative Evaluation?

The purpose of a Summative Evaluation is to assess the project's success. Summative Evaluation takes place after ultimate modifications and changes have been made, after the project is stabilized and after the impact of the project has had a chance to be realized. (Another term frequently used interchangeably with "Summative Evaluation" is "Impact Evaluation.") Summative Evaluation answers these basic questions:

The purpose of a Summative Evaluation is to assess the project's success.

- Was the project successful? What were its strengths and weaknesses?
- To what extent did the project or program meet the overall goal(s)?
- Did the participants benefit from the project? In what ways?
- What components were the most effective?
- Were the results worth the project's cost?
- Is this project replicable and transportable?

Summative Evaluation collects information about processes and outcomes. The evaluation is an external appraisal of worth, value or merit. Usually this type of evaluation is needed for decisionmaking. The decision alternatives may include the following: disseminate the intervention to other sites or agencies; continue funding; increase the funding; continue on probationary status; or discontinue.

Summative Evaluation informs decisionmakers about whether the activities and strategies were successful in helping the project and/or its participants reach their goals. This evaluation also describes the extent to which each goal was attained. Sample Summative Evaluation questions for a project like AMP could include the following:

- Did the majority of the undergraduate students in the project graduate with majors in mathematics, engineering or science?
- What proportion of graduates pursued their education until they received doctorates in mathematics, engineering, or science?
- Which elements or combinations of elements (mentoring, counseling, tutoring, or financial support) were most effective in retaining students in the SEM pipeline?

An important idea to keep in mind in conducting a Summative Evaluation is what has been called "unanticipated outcomes." These are findings that come to light during data collection or data analyses that were never anticipated when the study was first designed. An example of an unanticipated finding comes from a study that started out to look at whether or not school buses should have seat belts. This study also looked at the cost of purchasing new buses that had seat belts, versus retrofitting old models. This study, prompted by a desire to assure the safety of students, was ultimately unable to reach definitive conclusions regarding the utility of seat belts from the data available. Along the way, however, it was found that buses manufactured before a certain date were missing other safety features and the safety of the transportation system could be greatly enhanced by replacing buses purchased before this date. This unanticipated outcome became the basis for significant changes in the system's transportation policy.

An important idea to keep in mind while conducting a Summative Evaluation is to be vigilant of "unanticipated outcomes."

Summary

Evaluations can serve many different needs and provide critical data for decision-making at all steps of project development and implementation. Although some people feel that evaluation is an act that is done to a project, if done well, an evaluation is really done for the project.

It is important to remember that evaluation is not a

single thing, it is a process. When done well, evaluation can help inform the managers of the project as it progresses, can serve to clarify goals and objectives, and can provide important information on what is, or is not, working, and why.

This chapter has presented information to help Principal Investigators and project evaluators understand the various types of evaluation, the different stages in the evaluation process at which they occur, and the different kinds of information they provide. To summarize this information, a restatement of the important issues has been developed (see pages 12 and 13) to serve as a "shorthand" guide. For additional discussion of the various types of evaluation prototypes see Rossi and Freeman (1993). Chapter Five in this Handbook presents some additional examples of evaluations that further illustrate these roles and their differences.

REFERENCES

Rossi, P. H. & Freeman, H. E. (1993). *Evaluation:—A Systematic Approach (5th Edition)*. Newbury Park, CA: Sage.

Joint Committee on Standards for Educational Evaluation. (1981). *Standards for Evaluation of Educational Programs, Projects, and Materials*. New York, NY: McGraw-Hill.

Overview of Evaluation Prototypes

Planning Evaluation:

A Planning Evaluation assesses the understanding of project goals, objectives, strategies and timelines.

It addresses the following types of questions:

- Why was the project developed? What is the problem or need it is attempting to address?
- Who are the stakeholders? Who are the people involved in the project? Who are the people interested in the project who may not be involved?
- What do the stakeholders want to know? What questions are most important to which stakeholders? What questions are secondary in importance? Where do concerns coincide? Where are they in conflict?
- Who are the participants to be served?
- What are the activities and strategies that will involve the participants? What is the intervention? How will participants benefit? What are the expected outcomes?
- Where will the program be located (educational level, geographical area)?
- How many months of the school year or calendar year will the program operate? When will the program begin and end?
- How much does it cost? What is the budget for the program? What human, material, and institutional resources are needed? How much is needed for evaluation? for dissemination?
- What are the measurable outcomes? What is the expected impact of the project in the short run? the longer run?
- What arrangements have been made for data collection? What are the understandings regarding record keeping, responding to surveys, and participation in testing?

Formative Evaluation

A Formative Evaluation assesses ongoing project activities. It consists of two types: Implementation Evaluation and Progress Evaluation.

Implementation Evaluation

An Implementation Evaluation assesses whether the project is being conducted as planned. It addresses the following types of questions:

- Were the appropriate participants selected and involved in the planned activities?
- Do the activities and strategies match those described in the plan? If not, are the changes in activities justified and described?
- Were the appropriate staff members hired, and trained, and are they working in accordance with the proposed plan? Were the appropriate materials and equipment obtained?
- Were activities conducted according to the proposed timeline? by appropriate personnel?
- Was a management plan developed and followed?

Progress Evaluation

A Progress Evaluation assesses the progress made by the participants in meeting the project goals. It addresses the following types of questions:

- Are the participants moving toward the anticipated goals of the project?
- Which of the activities and strategies are aiding the participants to move toward the goals?

Summative Evaluation

A Summative Evaluation assesses project success—the extent to which the completed project has met its goals. It addresses the following types of questions:

- Was the project successful?
- Did the project meet the overall goal(s)?
- Did the participants benefit from the project ?
- What components were the most effective?
- Were the results worth the project's cost?
- Is this project replicable and transportable?

CHAPTER TWO: THE EVALUATION PROCESS — AN OVERVIEW

In the preceding chapter, we outlined the types of evaluations that Principal Investigators and Project Evaluators may want to carry out. In this chapter we talk further about how to carry out an evaluation, expanding, in particular, on two types of studies, Formative and Summative. In the sections that follow we provide an orientation to some of the basic language of evaluation, as well as some hints about technical, practical, and political issues that should be kept in mind in conducting evaluation efforts. Our goal is to capture a snapshot of the various pieces that make up the evaluation process from planning to report writing.

This overview is limited to topics pertinent to content and technique and does not cover other practical issues, such as budget planning, time tables, etc. Information on such topics can be found in the *Project Evaluation Kit* described in Chapter 8 (Bibliography). We have also limited the discussion to the types of evaluation most frequently carried out for NSF.

What are the Steps in Conducting a Formative or Summative Evaluation?

Whether they are Summative or Formative, evaluations can be thought of as having five phases:

- Develop evaluation questions
- Match questions with appropriate information-gathering techniques
- Collect data
- Analyze data
- Provide information to interested audiences.

All five phases are critical for provision of useful information. If the information gathered is not perceived as valuable or useful (the wrong questions were asked) or the information is not credible or feasible (the wrong techniques were used), or the report is presented too late or is written inappropriately, then the evaluation will not contribute to the decisionmaking process.

In the sections below we provide an overview of each of these phases, describing the activities that need to

take place in each. This overview is intended to provide a basic understanding of conducting an evaluation from start to finish. In Chapters Three and Four we provide greater detail in selected areas. A checklist summarizing the complete process is presented at the end of this chapter.

How Do You Develop Evaluation Questions?

The development of evaluation questions consists of several steps:

- Clarify goals and objectives of the evaluation
- Identify and involve key stakeholders and audiences
- Describe the intervention to be evaluated
- Formulate potential evaluation questions of interest to all stakeholders and audiences
- Determine resources available
- Prioritize and eliminate questions.

Getting started right can have a major impact on the progress of an evaluation all along the way.

Although it may sound trivial, at the outset of an evaluation it is important to describe the project or intervention briefly and clarify goals and objectives of the evaluation. Getting started right can have a major impact on the progress of an evaluation all along the way. Patton (1990) suggests considering the following questions in developing an evaluation approach:

- Who is the information for and who will use the findings?
- What kinds of information are needed?
- How is the information to be used? For what purpose is evaluation being conducted?
- When is the information needed?
- What resources are available to conduct the evaluation?
- Given the answers to the preceding questions, what methods are appropriate?

A critical component of clarifying goals and objectives is the identification of the evaluation's focus. Is the

evaluation to be Formative, looking at, for example, whether or not a teacher enhancement activity has been implemented as planned? Or is it Summative, looking at the impact of the program on teaching practices and, ultimately, student learning?

Equally important is the identification of either stakeholders in the project or potential audiences for the evaluation information. In all projects, multiple audiences are likely to be involved. Being clear about your audience is very important as different audiences will have different information needs. For example, the kinds of information needed by those who are concerned about the day-to-day operations of a project will be very different from those needed by policymakers who may be dealing with more long-term issues or who have to make funding decisions.

The next step is a goal-oriented description of the project including the rationale given for its existence and its goals and objectives as seen by the stakeholders. The essence of the intervention should also be documented: where it is situated, who is involved, how it is managed, and how much it costs. An in-depth understanding of the intervention is usually necessary to determine the full range of evaluation questions. This type of goal-centered description is often a significant part of the evaluation effort.

It is critical to identify the major stakeholders, their questions, and their needs for information.

After the purpose and stakeholders are identified and the project is described, specific questions about the project should be formulated. The process of identifying target audiences and formulating potential questions will usually result in many more questions than can be addressed in a single evaluation effort. This comprehensive look at potential questions, however, makes all of the possibilities explicit to the planners of the evaluation and allows them to make an informed choice among evaluation questions. Each potential question should be considered for inclusion on the basis of the following criteria:

- Who would use the information
- Whether the answer to the question would provide information not now available
- Whether information is important to a major group or several stakeholders
- Whether information would be of continuing interest

- Whether it would be possible to obtain the information, given financial and human resources.
- Whether the time span required to obtain the information would meet the needs of decision makers.

These criteria determine the relevance of each potential question.

A general guideline is for evaluations to be funded at about 5-10 percent of the total project cost.

The final selection of questions depends heavily on the resources available. Some evaluation activities are more costly than others. For example, it may be that the only way to answer the question: "How has a project designed to enhance teachers' classroom activities affected classroom practices?" is through extensive classroom observations, an expensive and time-consuming technique. If sufficient funds are not available to carry out observations, it may be necessary to reduce the sample or use a different data collection method such as a survey. A general guideline is to allocate 5-10 percent of project costs for the evaluation of large-scale projects (those exceeding $100,000); for smaller projects, the percentage may need to be higher to meet minimum costs of fielding evaluation activities.

How Do You Determine the Information-Gathering Techniques?

The next stage is the determination of the appropriate information-gathering techniques, including several steps:

- Select a general methodological approach
- Determine what sources of data would provide the information needed and assess the feasibility of the alternatives
- Select data collection techniques that would gather the desired information from the identified sources
- Develop a design matrix.

After the evaluation questions have been formulated, the most appropriate methods for obtaining answers must be chosen. In determining what approach to use, some initial questions need to be answered. First, is it better to do case studies, exploring the experiences

of a small number of participants in depth or is it better to use a survey approach? In the latter case, do you need to survey all participants or can you select a sample? Do you want to look only at what happens to project participants or do you want to compare the experiences of participants with those of some appropriately selected comparison group of nonparticipants? How you answer some of these questions will affect the kinds of conclusions you can draw from your study. Rigorous, "controlled" designs are not always needed for Formative (Process) Evaluations, although they are always the preferred design. But Summative Evaluations, or Impact Assessments, gain a great deal from being based on experimental or quasi-experimental designs. For more information on this and the implications of different evaluation designs, see Cook and Campbell (1979).

Next you need to determine the kinds of data you want to use. Some alternatives are listed in Exhibit 2. Which one or ones to use depends on a number of factors, including the questions, the timeline and the resources available. Another factor to take into account is the technical skill level of the evaluator or evaluation team. Some of the techniques require more skills than others to design and analyze. If you are limited in your evaluation resources, it is best to stick to the simpler approaches. For example, observational techniques can produce a rich database which, analyzed properly, can be highly informative. The trick here is to design instruments which are either suitable for statistical analysis, or for other analytic strategies which have been developed for case study evidence (Yin, 1989). In the absence of careful advance planning for the analysis, many an evaluator has wound up with a massive investment (both in time and in money) of data collected via observation that elude reasonable analysis.

If you are limited in your evaluation resources, it is best to stick to the simpler approaches.

Finally, you need to decide on the appropriate mix of data collection techniques, including both quantitative and qualitative approaches.

In a broad sense, quantitative data can be defined as any data that can be represented numerically, whereas qualitative data are more frequently expressed through narrative description. Quantitative data also are useful in measuring the reactions or skills of large groups of people on a limited set of questions, whereas qualitative data provide in-depth information on a smaller number of cases (Patton, 1990). These distinc-

Exhibit 2

Sources and Techniques for Collecting Evaluation Information

I. Data Collected Directly From Individuals Identified as Sources of Information

A. Self-Reports: (from participants and control group members)

1. Diaries or Anecdotal Accounts
2. Checklists or Inventories
3. Rating Scales
4. Semantic Differentials
5. Questionnaires
6. Interviews
7. Written Responses to Requests for Information (for example, letters)
8. Sociometric Devices
9. Projective Techniques

B. Products from participants:

1. Tests
    a. Supplied answer (essay, completion, short response, and problem-solving)
    b. Selected answer (multiple-choice, true-false, matching, and ranking)
2. Samples of Work

II. Data Collected by an Independent Observer

A. Written Accounts

B. Observation Forms:

1. Observation Schedules
2. Rating Scales
3. Checklists and Inventories

III. Data Collected by a Mechanical Device

A. Audiotape

B. Videotape

C. Time-Lapse Photographs

D. Other Devices:

1. Graphic Recordings of Performance Skills
2. Computer Collation of Student Responses

IV. Data Collected by Use of Unobtrusive Measures

V. Data Collected from Existing Information Resources

A. Review of Public Documents (proposals, reports, course outlines, etc.)

B. Review of Institutional or Group Files (files of student records, fiscal resources, minutes of meetings)

C. Review of Personal Files (correspondence files of individuals reviewed by permission)

D. Review of Existing Databases (statewide testing program results)

From: *Education Evaluation: Alternative Approaches and Practical Guidelines.* by Blaine R. Worthen and James R. Sanders. Copyright ©1987 by Longman Publishing Group.

tions are not, however, absolute. Rather, they can be thought of as representing two ends of a continuum rather than two discrete categories. Furthermore, in some instances qualitative data can be transformed into quantitative data using judgmental coding (for example grouping statements or themes into larger broad categories and obtaining frequencies). Conversely, well-designed quantitative studies will allow for qualitative inputs.

Both types of data can provide bases for decisionmaking; both should be considered in planning an evaluation. And evaluations frequently use a mix of techniques in any one study. Further details on data collection and analysis techniques and the pros and cons of different alternatives are presented in Chapter Three of this Handbook.

Once these decisions are made it is very helpful to summarize them in a "design matrix." Although there is no hard and fast rule, a design matrix usually includes the following elements:

- General evaluation questions
- Evaluation subquestions
- Variables to be examined and instruments/ approaches for gathering the data
- Respondents
- Data collection schedule.

Exhibit 3 presents an example of a design matrix for a study of the effects of a teacher enhancement program.

How Do You Conduct Data Collection?

Once the appropriate information-gathering techniques have been determined, the information must be gathered. Both technical and political issues need to be addressed. The technical issues are discussed in Chapter Three. The political factors to be kept in mind are presented below:

- Obtain necessary clearances and permission
- Consider the needs and sensitivities of the respondents

Exhibit 3: Summary of the Design for a Study of Project SUCCEED

Question 1: Did project SUCCEED change teachers' mathematics instructional practices?			
Subquestion	**Data Collection Approach**	**Respondents**	**Schedule**
1a. Did teachers use different materials?	Questionnaire	Teachers	Pre/post training
	Observation	Supervisors NA	3 x during year
1b. Did teachers change testing practices?	Questionnaire	Teachers	End of year
1c. Was cooperative learning increased?	Questionnaire	Teachers	End of year
	Observation	NA	3 x during year
Question 2: What impact did project SUCCEED have on teachers' use of planning time?			
Subquestion	**Data Collection Approach**	**Respondents**	**Schedule**
2a. Did teachers spend more time planning for instruction?	Questionnaire	Teachers	End of year
2b. Did teachers develop lesson plans reflecting new approaches?	Questionnaire	Teachers	End of year
	Review of plans	NA	3 x during year
2c. Did teachers use the sample lessons as models?	Questionnaire	Teachers	End of year
	Review of plans	NA	3 x during year

- Make sure your data collectors are adequately trained and will operate in an objective, unbiased style

- Cause as little disruption as possible to the ongoing effort.

First, before data are collected, the necessary clearances and permission must be obtained. Many groups, especially school systems, have a set of established procedures for gaining clearance to collect data on students, teachers, or projects. This may include who is to receive/review a copy of the report, restrictions on when data can be collected, or procedures to safeguard the privacy of students or teachers. Find out what these are and address them as early as possible, preferably as part of the initial proposal development. When seeking cooperation, it is always helpful to offer to provide feedback to the participants on what is learned. Personal feedback or a workshop in which findings can be discussed is frequently looked upon favorably. If this is too time-consuming, a copy of the report or executive summary may well do. The main idea here is to provide incentives for people or organizations to take the time to participate in your evaluation.

The main idea here is to provide incentives for people or organizations to take the time to participate in your evaluation.

Second, the needs of participants must be considered. Being part of an evaluation can be very threatening to participants. Participants should be told clearly and honestly why the data are being collected and the use to which the results will be put. On most survey type studies, assurances are given and honored that no personal repercussions will result from information presented to the evaluator and, if at all possible, individuals and their responses will not be publicly associated in any report. This guarantee of anonymity frequently makes the difference between a cooperative and a recalcitrant respondent. There may, however, be some cases when identification of the respondent is deemed necessary, perhaps to enforce the credibility of an assertion. In such cases, the evaluator should seek informed consent before including such information. Informed consent may also be advisable where a sensitive comment is reported which could be identified with a given respondent, despite the fact that the report itself includes no names. Common sense is the key here.

Third, data collectors must be carefully trained and supervised, especially where multiple data collectors are used. They must be trained to see things in the same way, to ask the same questions, to use the same

prompts. Periodic checks need to be carried out to make sure that well-trained data collectors do not "drift" away from the prescribed procedures over time. (More details on training of data collectors are presented in Chapter Three.)

Checks need to be carried out to make sure that well-trained data collectors do not "drift" away from the prescribed procedures.

In addition, it is important to guard against possible distortion of data due to well intentioned but inappropriate "coaching" of respondents —an error frequently made by inexperienced or overly enthusiastic staff. They must be warned against providing value-laden feedback to respondents or to engage in discussions that might well bias the results. One difficult but important task is understanding one's own biases and making sure that they do not interfere with the work at hand. This is a problem all too often encountered when dealing with volunteer data collectors, such as parents in a school or teachers in a center. They volunteer because they are interested in, advocates for, or critics of, the project that is being evaluated. Unfortunately, the data they produce may reflect their own perceptions of the project, as much or more than that of the respondents, unless careful training is undertaken to avoid this "pollution." Bias or perceived bias may compromise the credibility of the findings and the ultimate use to which they are put. An excellent source of information on these issues is the section on accuracy standards in Standards for Evaluation of Educational Programs, Projects and Materials (Joint Committee on Standards for Educational Evaluation, 1981).

Finally, the data should be gathered causing as little disruption as possible. Among other things, this means being sensitive to the schedules of the people or the project, as well as the schedule of the evaluation itself. It also may mean changing approaches as situations come up. For example, instead of asking a respondent to provide data on the characteristics of project participants—a task that may require considerable time on the part of the respondent to pull the data together and develop summary statistics—the data collector may have to work from raw data, applications, monthly reports, etc. and personally do the compilation.

How Do You Analyze the Data?

Once the data are collected they must be analyzed and interpreted. The steps to be followed in preparing the data for analysis and interpretation differ, depending on the type of data. The interpretation of qualitative data may in some cases be limited to descriptive

narratives, but other qualitative data may lend themselves to systematic analyses through the use of quantitative approaches such as thematic coding or content analysis. Analysis includes several steps:

- Check the raw data and prepare data for analysis
- Conduct initial analysis based on the evaluation plan
- Conduct additional analyses based on the initial results
- Integrate and synthesize findings.

The first step in quantitative data analysis is the checking of data for responses that may be out of line or unlikely. Such instances include: selecting more than one answer when only one can be selected; always choosing the third alternative on a multiple-choice test of science concepts; reporting allocations of time that add up to more than 100 percent; inconsistent answers, etc. Where such problematic responses are found, it is frequently necessary to eliminate the item or items from the data to be analyzed.

Data analysis is an iterative process.

After this is done, the data are prepared for computer analysis; usually this involves coding and entering (keying) the data with verification and quality control procedures in place.

The next step is to carry out the data analysis specified in the evaluation plan. While new information gained as the evaluation evolves may well cause some analyses to be added or subtracted, it is a good idea to start with the set of analyses that seemed to be of interest originally. For the analysis of both qualitative and quantitative data there are statistical programs currently available on easily accessible software that make the data analysis task considerably easier today than it was 25 years ago. These should be used. Analysts still need to be careful, however, that the data sets they are using meet the assumptions of the technique being used. For example, in the analysis of quantitative data, different approaches may be used to analyze continuous data as opposed to categorical data. Using an incorrect technique can result in invalidation of the whole evaluation project. (See Chapter Three for more discussion of alternative analytic techniques.)

It is very likely that the initial analyses will raise as many questions as they answer. The next step, there-

fore, is conducting a second set of analyses to address these further questions. If, for example, the first analysis looked at overall teacher performance, a second analysis might want to subdivide the total group into subunits of particular interest—i.e., more experienced versus less experienced teachers—and examine whether any significant differences were found between them. These reanalysis cycles can go through several iterations as emerging patterns of data suggest other interesting avenues to explore. Sometimes the most intriguing of these are results which emerge from the data; ones that were not anticipated or looked for. In the end, it becomes a matter of balancing the time and money available, against the inquisitive spirit, in deciding when the analysis task has been completed.

The final task is to choose the analyses to be presented, to integrate the separate analyses into overall pictures, and to develop conclusions regarding what the data show. Sometimes this integration of findings becomes very challenging as the different data sources do not yield completely consistent results. While it is always preferable to produce a report that is able to reconcile differences and explain apparent contradictions, sometimes the findings must simply be allowed to stand as they are, unresolved and thought-provoking.

How Do You Communicate Evaluation Results?

The final stage of the Project Evaluation is reporting what has been found. While reporting can be thought of as simply creating a written document, successful reporting rests on giving careful thought to the creation and presentation of the information. In fact, while funding agencies like NSF require a written report, many projects use additional strategies for communicating evaluation findings to other audiences.

The communication of evaluation findings involves several steps:

- Provide information to the targeted audiences
- Customize reports and other presentations to make them compelling
- Deliver reports and other presentations in time to be useful.

Providing the evaluation information should not present a problem if the evaluation has been successful so far, and if some simple steps are followed. Again, special attention should be given to the stakeholders and the constructive part they can play. The specification of questions and selection of data-gathering techniques should have already involved the stakeholders so that the information should be relevant and important to them. By also involving the stakeholders at the end of the study, the utility and probable attention given to the evaluation findings are sure to be increased. One way of accomplishing this is through a pre-release review of the report with selected stakeholder representatives. Such a session provides an important opportunity for discussion of the findings, for resolving any final issues that may arise and for setting the stage for the next steps to be taken as a result of the successes and failures that the data may show.

By involving the stakeholders at the end of the study, the utility and attention given to the evaluation findings are sure to be increased.

Second, the information must be delivered when it is needed. Sometimes there is leeway in when the information will be used; but the time of decision-making is often fixed, and information that arrives too late is useless. There is nothing so frustrating to a Principal Investigator than being told by a funding agency or community group:

"Oh, I wish I had known that two months ago! That's when I had to make some decisions about the projects we were going to support next year."

Our earlier discussion stressed the importance of agreeing up front what is needed and when the needs must be met. As the evaluation is carried out, the importance of meeting the agreed-upon time schedule must be kept in mind.

Finally, the information needs to be provided in a manner and style that is appropriate, appealing, and compelling to the person being informed. For example, a detailed numerical table with statistical test results might not be the best way to provide a school board member with achievement data on students. Different reports may have to be provided for different audiences. And, it may well be that a written report is not even the preferred alternative. While most evaluations will include some written product, other alternatives are becoming increasingly popular.

It should be noted that while discussions of commu-

nicating study results generally stop at the point of presenting a final report of findings, there are important additional steps that should be considered. Where a new product or practice turns out to be successful, as determined by a careful evaluation, dissemination is an important next step. This topic is covered in a separate NSF publication.

Summary

There are several phases to conducting and implementing an evaluation. No one stage is more important than the rest. And, as can be seen from the discussion of the role of the stakeholders in both the first step—developing questions—and the last—provision of information—the groundwork laid at the earliest stages can have important implications for the success of the evaluation in the long run.

While there are technical skills needed to do an evaluation, there is also "common sense" involved.

Evaluation isn't easy, but there also is very little mystery about the steps that need to be taken and the activities that need to be carried out. While there certainly are technical skills needed to do an evaluation that is helpful and credible—and that is why trained evaluators are important—there is also a lot of "common sense" involved. Sound advice is to blend these two factors—technical skills and common sense. In the best evaluations, both of these inevitably exist.

REFERENCES

Cook, T. D. & Campbell, D. T. (1979). *Quasi-experimentation: Design and Analysis Issues or Field.* Settings, Chicago, IL: Rand McNally.

Joint Committee on Standards for Educational Evaluation. (1981). *Standards for Evaluation of Educational Programs, Projects, and Materials.* New York, NY: McGraw-Hill.

Patton, M. Q. (1990). *Qualitative Evaluation and Research Methods.* Newbury Park, CA: Sage.

Worthen, B. & Sanders. J. (1987). *Educational Evaluation: Alternative Approaches and Practical Guidelines.* New York: Longman, Inc.

Yin, R. (1989). *Case Study Research.* Newbury Park, CA: Sage.

Tips for Conducting an Evaluation

1. Develop Evaluation Questions

- Clarify goals and objectives of the evaluation.
- Identify and involve key stakeholders and audiences.
- Describe the intervention to be evaluated.
- Formulate potential evaluation questions of interest to all stakeholders and audiences.
- Determine resources available.
- Prioritize and eliminate questions.

2. Match Questions with Appropriate Information-Gathering Techniques

- Select a general methodological approach.
- Determine what sources of data would provide the information needed.
- Select data collection techniques that would gather the desired information from the identified sources.

3. Collect Data

- Obtain the necessary clearances and permission.
- Consider the needs and sensitivities of the respondents.
- Make sure data collectors are adequately trained and will operate in an objective, unbiased manner.
- Cause as little disruption as possible to the ongoing effort.

4. Analyze Data

- Check raw data and prepare data for analysis.
- Conduct initial analysis based on the evaluation plan.
- Conduct additional analyses based on the initial results.
- Integrate and synthesize findings.

5. Provide Information to Interested Audiences

- Provide information to the targeted audiences.
- Deliver reports and other presentations in time to be useful.
- Customize reports and other presentations.

CHAPTER FOUR: REPORTING

The product of an evaluation is almost always a formal report. While the report frequently may be supplemented by other forms of oral communication—overheads, conference presentations, workshops—a formal written report is standard for NSF. Depending on exactly who the audience is, a specific report may vary in format, length, and level of technical discussion. For example, a report to a Board of Education will be far more concise, and less technical, than a report to a professional association or a funding agency.

In this chapter, we discuss the development of a formal report for an agency like the National Science Foundation. The specific type of report on which we focus is one that would be the product of an experimental or quasi-experimental design. For details on developing reports for other methodologies, specifically, case studies, see Yin, (1989).

What are the Components of a Formal Report?

Most reports include five major sections. The major sections are:

- Background
- Evaluation Study Questions
- Sample, Data Collection, Instrumentation
- Findings
- Conclusions (and recommendations).

The Background Section

The background section includes and describes the following: (1) the problem or needs addressed; (2) the stakeholders and their information needs; (3) the participants; (4) the project's objectives; (5) the activities and components; (6) location and planned longevity of the project; (7) the resources used to implement the project; and (8) the project's expected measurable outcomes.

Notable constraints that existed in what the evaluation was able to do are also pointed out in this section.

For example, it may be important to point out that the conclusions are limited by the fact that no appropriate comparison group was available or that only the short term effects of program participation could be examined.

Evaluation Study Questions

The evaluation is based on the need for specific information; stakeholders such as Congress, NSF-funded program and project directors, and the participants, have distinct needs. There are many questions to be asked about a project. However, all of these questions cannot be answered at one time. This section of the report describes the questions that the study addressed. As relevant, it also points out some important questions that could not be addressed due to factors such as time, resources, or inadequacy of available data gathering techniques.

Evaluation Procedures

This section of the report describes the groups that participated in the evaluation study. For quantitative studies it describes who these groups were and how the particular sample of respondents included in the study was selected from the total population available, if sampling was used. Important points noted are how representative the sample was of the total population; whether the sample volunteered (self-selected) or was chosen using some sampling strategy by the evaluator; and whether or not any comparison or control groups were included.

This section also describes the types of data collected and the instruments used for the data collection activities. For example, they could be:

- Quantitative data for identified critical indicators, e.g., grades for specific subjects, Grade Point Averages (GPA's)
- Ratings obtained in questionnaires and interviews designed for project directors, students, faculty, and graduate students
- Descriptions of classroom activities from observations of key instructional components of the project
- Examinations of extant data records, e.g., letters, planning papers, and budgets.

It is helpful at the end of this section to include a "matrix" or table which summarizes the evaluation questions, the variables, the data-gathering approaches, the respondents, and the data collection schedule.

Data Analysis

This section describes the techniques used to analyze the data collected above. It describes the various stages of analysis that were implemented and the checks that were carried out to make sure that the data were free of as many confounding factors as possible. Frequently, this section contains a discussion of the techniques used to make sure that the sample of participants that actually participated in the study was, in fact, representative of the groups from which they came. (That is, there is sometimes an important distinction between the characteristics of the sample that was selected for participation in the evaluation study and the characteristics of those who actually participated—returned questionnaires, attended focus groups, etc.)

Again, a summary matrix is a very useful illustrative tool.

Findings

This section presents the results of the analyses described previously. The findings are usually organized in terms of the questions presented in the section on Evaluation Study Questions. Each question is addressed, regardless of whether or not a satisfactory answer is provided. It is just as important to point out where the data are inconclusive, as where the data provide a positive or negative answer to an evaluation question. Visuals such as tables and graphical displays are an appropriate complement to the narrative discussion.

While the discussion in the findings section usually focuses most heavily on quantitative information, qualitative information may also be included. In fact, including both can turn a rather "dry" discussion of results into a more meaningful communication of study findings. An easy way to do this is to include quotes from the project participants which help to illustrate the point being made.

At the end of the findings section, it is helpful to have a summary that presents the major conclusions. Here "major" is defined in terms of both the priority of the

question in the evaluation and the strength of the finding from the study. For example, in a Summative Evaluation, the summary of findings would always include a statement of what was learned with regard to outcomes, regardless of whether or not the data were conclusive.

Conclusions (and Recommendations)

The conclusions section reports the findings with more broad-based and summative statements. These statements must relate to the findings of the project's evaluation questions and to the goals of the overall program. Sometimes the conclusions section goes a step further and includes recommendations either for the Foundation or for others undertaking projects similar in goals, focus, and scope. Care must be taken to base any recommendations solely on robust findings and not on anecdotal evidence, no matter how persuasive.

Other Sections

In addition to these six major sections, formal reports also include one or more summary sections. These would be:

- An Abstract—a summary of the study and its findings presented in approximately one-half a page of text.
- An Executive Summary—a summary which may be as long as four pages that provides an overview of the evaluation, its findings, and implications. Sometimes the Executive Summary also serves an a non-technical digest of the evaluation report.

How Do You Develop an Evaluation Report?

Although we usually think about report writing as the last step in an evaluation study, a good deal of the work actually can and does take place before the project is completed. The "Background" section, for example, can be based largely on the original proposal. While there may be some events that cause minor differences between the study as planned and the study as implemented, the large majority of information such as research background, the problem addressed, the stakeholders and the project's goals, will remain essentially the same.

If you have developed a written evaluation design, the

material in this design can be used for the sections on "Evaluation Study Questions" and "Sample, Data Collection, Instrumentation." The "Data Analysis" section is frequently an updated version of what was initially proposed. However, as we noted in Chapter Two, data analysis can take on a life of its own, as new ideas emerge when data are explored; the final data analysis may be far different than what was initially envisioned.

The "Findings" and "Conclusions" sections are the major new sections to be written at the end of an evaluation study. These may present somewhat of a challenge because of the need to balance comprehensiveness with clarity, and rigorous, deductive thinking with intuitive leaps.

One of the errors frequently made in developing a "Findings" section is what we might call the attitude of "I analyzed it, so I am going to report it." That is, evaluators may feel compelled to report on analyses that appeared fruitful, but ultimately resulted in little information of interest. In most cases, it is sufficient to note these analyses were conducted and that the results were inconclusive. Presentation of tables showing that no differences occurred or no patterns emerged, is probably not a good idea unless there is a strong conceptual or political reason for doing so. Even in the latter case, it is prudent to note the lack of findings in the text and to provide the back-up evidence in appendices or some technical supplement.

One tip to follow when writing these last sections is to ask colleagues to review what you have written and provide feedback before the report reaches its final form. Your colleagues can assist in assessing the clarity and completeness of what you have written, as well as provide another set of eyes to examine your arguments and, possibly, challenge your interpretations. It is sometimes very hard to get enough distance from your own analyses after you have been immersed in them. Ask a colleague for help and return that favor in the future.

What Might a Sample Report Look Like?

In the pages that follow, we present sample sections from an evaluation report. The sections have been created to illustrate further the kinds of information included in each section. This report is a progress report developed after the first year of funding of a project that will ultimately continue for a total of 5 years.

Report of the Higher Education University Alliances for Minority Participation (AMP) Project

Background

Overview of the AMP Program

The Alliances for Minority Participation (AMP) Program is funded by the National Science Foundation's Human Resource Development division, part of the minority student development initiative. The program was developed in response to concerns raised by the low number of underrepresented minority students who successfully completed science and engineering baccalaureate degree programs. A major goal of the AMP Program is to increase substantially the pool of interested and academically qualified underrepresented minority students who go on for graduate study in these fields. Students eligible to participate in this program are United States citizens or legal residents in undergraduate colleges and universities who are African American, American Indian/Alaskan Native, or Hispanic.

AMP Program objectives are listed below:

- Establish partnerships among community colleges, colleges and universities, school systems, Federal/state/local agencies, major national Science, Engineering and Mathematics (SEM) laboratories and centers, industry, private foundations, and SEM professional organizations.
- Provide activities that facilitate the transition and advancement of minority students through one or more critical decision points during SEM education—high school to college, 2-year to 4-year college, undergraduate to graduate school, or college to the workplace.
- Achieve a demonstrated increase in the number of underrepresented minority students receiving undergraduate SEM degrees.
- Demonstrate the involvement and commitment of SEM departments and faculty in the design and implementation of improvements of SEM undergraduate education.
- Demonstrate the existence of an infrastructure and management plan for ensuring long-term continuance of AMP or similar activities among the participating organizations and institutions.
- Identify for evaluative purposes the critical data elements associated with demonstrating the increases of undergraduate and graduate students in SEM programs.

Project Description

The "Higher Education University" AMP project was funded initially for 5 years, renewable each year. The project operated during 1991-92 from September to June and during the 1992 summer session from July to August. According to the project plan, some students participated during both periods, while others participated in either the

academic year or the summer session. The "Higher Education University" AMP project included a coalition of colleges, universities, K-12 schools and nonprofit organizations.

Students were involved in the following core activities:

- Summer research using the skills of science, engineering, and/or mathematics
- Travel to attend professional meetings for scientists, engineers, and mathematicians
- Attendance at programs to hear presentations by special speakers on subjects related to science, engineering or mathematics
- Enrollment in the Summer Bridge Program to improve and enrich SEM skills
- Participation in peer study groups to improve skills in science and/or mathematics.
- A Black History program that was added to enhance African American students' self-esteem.

The project was fully funded for $1 million per year. This funding provided for project managers, faculty/teachers, graduate students, support personnel, development of a project database, student financial support, and the implementation of the previously described activities.

The Purpose of the Project and Its Evaluation

The goal of the project was to provide appropriate activities and support to students involved in the AMP project to improve and enrich their science, engineering and mathematics skills so that their interest and retention in the SEM pipeline continues through undergraduate to graduate school, and eventually to the Ph.D. The initial evaluation of the project focused on identifying those activities that successfully met the AMP objectives and on reporting student outcomes related to the success of the activities. On an annual basis, the evaluation will identify which of the project activities need to be modified or deleted prior to the project's Summative Evaluation.

"Higher Education University" AMP Project objectives are listed below:

- Establish and maintain a partnership among community colleges, colleges and universities, school systems, and industry
- Provide activities that facilitate the transition and advancement of minority students through two critical decision points during SEM education— 2-year to 4-year college and undergraduate to graduate school
- Retain 95 percent of the AMP students in SEM courses
- Increase the number of underrepresented minority students in SEM courses each year by 10 percent

- Increase by 25 percent each year the number of minority students receiving undergraduate SEM degrees
- Demonstrate the involvement and commitment of SEM departments and faculty in the design and implementation of improvements of SEM undergraduate education
- Demonstrate the existence of an infrastructure and management plan for ensuring long-term continuance of AMP or similar activities among the participating organizations and institutions
- Identify which components and activities helped recruit, retain and increase the number undergraduate and graduate students in SEM programs.

Evaluation Questions

The evaluation looked at a broad range of questions related to both the project's implementation and its success. Specifically, the evaluation addresses the following questions:

- Did the AMP project result in the establishment of adequate partnerships?
- What was the impact of the partnerships on promoting the AMP project's objectives?
- What activities were most successful in recruiting, retaining and increasing underrepresented minorities in science, engineering, and mathematics?
- What evidence is there that the project may successfully reach its long-term outcomes (e.g., SEM baccalaureate degree, acceptance into graduate school seeking a SEM degree)?
- How could the project be improved and/or changed to better serve the needs of underrepresented minority students who are enrolled in science, engineering, and mathematics courses?

Sample, Data Collection, Instrumentation

The primary sources for information about the AMP project came from the Annual Reports, the database of Minimum Obligatory Set (MOS) elements, project-focused questionnaires and interviews, and observations of the instructional components. The principal group for study was all AMP students in the project. Aggregates of grades by race, gender and class status, pass and fail records, GPAs and retention rates were the basis for analyzing the impact of the "Higher Education University" AMP project sites and components. These same data were used for tracking the longitudinal progress of the AMP project for SEM underrepresented minority students—that is, the students' movement towards SEM baccalaureate and graduate degrees. For comparative purposes, selected scholastic information for all SEM students by race, gender and class status were collected. Table 1 summarizes the evaluation design. Variables, measures, and samples (participants) are presented for each evaluation question.

Table 1: Summary of the Evaluation Design

Question 1: Did the AMP project result in the establishment of adequate partnerships?			
Subquestion	**Data Collection Approach**	**Respondents**	**Schedule**
1a. What partnerships were established?	Review of records Interviews	NA Principal Investigator	End of year End of year
1b. Were the partnerships established in a timely fashion?	Review of records Interviews	NA Principal Investigator	End of year
1c. Were the goals for the number and mix of partners achieved?	Comparison of proposal and data from 1a	NA	End of year
Question 2: What impact did the partnerships have?			
Subquestion	**Data Collection Approach**	**Respondents**	**Schedule**
2a. How effective were the partnerships?	Questionnaires	All staff, Selected students	End of year
2b. What were the most effective activities provided by them?	Questionnaires Observation	All staff NA	End of year Ongoing

Data Analysis

Descriptive statistics (frequencies, percentages, means, medians, standard deviations, etc.) were used to report the results of the evaluation. Also, these statistics were used to make comparisons among the ratings and statements from the various respondents. Tests of significance were computed to determine if there were real differences among certain quantitative data, that is, the grade point averages and grades for AMP students and all SEM students.

Table 2 summarizes the data analysis design. The measures, variables, and analyses are presented for each evaluation question.

Table 2: Summary of the Data Analysis Plan

Question 1: Did the AMP project result in the establishment of adequate partnerships?		
Subquestion	**Data Collection Approach**	**Analysis Plan**
1a. What partnerships were established?	Review of records Interviews	Descriptions Simple numerical tallies
1b. Were the partnerships established in a timely fashion?	Review of records Interviews	Frequency distribution of time of partnership establishment
1c. Were the goals for the number and mix of partners achieved?	Comparison of proposal and data from 1a	Matching of goals with achievements
Question 2: What impact did the partnerships have?		
Subquestion	**Data Collection Approach**	**Analysis Plan**
2a. How effective were the partnerships?	Questionnaires	Percentages selecting various ratings
2b. What were the most effective activities provided by them?	Questionnaires Observation	Percentages selecting various ratings Summaries of running records

Findings

Did the AMP project result in the establishment of adequate partnerships?

The total number of institutions and groups that formed coalitions with the AMP project was 25. The types of groups that formed the coalitions were: 2 colleges, 5 universities, 4 junior colleges, 3 school districts, and 11 businesses and community groups. The only group that was below the target was school districts in the area. The target was to have 6 school districts involved.

The activities of these coalitions were judged to be very helpful to the project. Over 80 percent of the staff responded "helpful" or "very helpful" (the highest ratings) to the following activities:

- Mentoring
- Extended on-site experiences
- Special training opportunities

etc.

Which components of the AMP project were the most successful in supporting and retaining AMP students?

The Summer Bridge Program was a very successful part of this project. This conclusion is based on ratings from the AMP students and the grades that they subsequently received in pre-calculus and calculus. (See Tables 3 and 4.)

The data are impressive. Over 50% of the Summer Bridge students received an "A" or a "B" in these classes. Equally important, the failure rate was low. Only 10% of the students failed pre-calculus and 15% failed calculus. These compare quite favorably with failure rates in the past which ranged from an average of 20% in pre-calculus to 35% in calculus.

A number of the comments about the Summer Bridge Program recognized the positive effects of the support systems which were provided.

A 16-year old female remarked:

"The Bridge Program gave me just that extra little boost that I needed to do well in my classes. I knew the basic material. I knew I knew the basic material. I was ready!"

This program was not without criticism, however. Several students pointed out that the schedule of classes, occurring as they did at mid-summer, prevented them from taking jobs that they needed. They recommended that the program be offered right after the end of the regular school year or right before the beginning of the next school year, so that a block of time would be available for employment.

Another component that received high ratings was the Peer Study Group with its opportunities for collaborative learning. (See Table 3.)

etc.

For a moderate number of students (35%), scholarship assistance was critical for them to remain in school.

etc.

Summary of Findings

1. There was a large number of coalitions formed between from the business and community groups. School districts met only half of the participation target.

2. The Summer Bridge Program, Peer Study Group, and scholarship assistance were AMP components, which were rated highly for supporting students to remain in the SEM pipeline.

etc.

Conclusions

For 1991-92, the "Higher Education University" AMP project met its objectives. The project established partnerships among community colleges, universities, school systems, business, and community groups. However, partnership with the school system fell short of their participation target by 50 percent.

The Summer Bridge Program, Peer Study Groups, and scholarship assistance were rated by the AMP students, faculty, and AMP staff as critical to retaining and increasing undergraduate and graduate students in SEM programs. However, in the Summer Bridge Program, there were some negative remarks about the scheduling of activities and recommendations were made for...

The Black History course received high ratings from the African American AMP students. Many of these students said that they were encouraged to succeed as SEM students after learning about successful role models in science and mathematics.

CHAPTER FIVE: EXAMPLES

Sometimes, providing an example is more worthwhile than a thousand words of advice. Because we believe that is the case with evaluation, we have provided a number of examples (both good and bad) of different evaluation efforts. These examples build on information from the previous chapters. The specific examples we present are fictitious, but are based on studies that could and have been done.

Example 1: Evaluation of an Inservice Program for Elementary Science Teachers

This example illustrates the:

- **Use of formative evaluation for measuring implementation**
- **Importance of involving a skilled evaluator early in the project**
- **Use of both qualitative and quantitative approaches**
- **Use of multi-informants and data collection techniques**
- **Ability to adapt a design based on new information.**

NSF funded the Center for Professional Enhancement for a 3 year program of in-service education for elementary science teachers. The purpose of the project was to introduce the teachers to some of the new approaches to elementary science instruction and to assist them in applying these techniques in their classrooms. Teachers were selected for participation based on the nomination of a supervisor (usually a principal), their written essay, and a commitment both to attend regular and summer sessions and try out the new approaches in their classrooms.

The program consisted of 1 year of training with follow-up workshops after the first year. A mixture of teaching strategies was used. These included: lecture sessions, hands-on experiences in using techniques identified as being promising, visits to classrooms taught by model teachers, and peer coaching.

Before the project was even funded, the Principal investigator hired an evaluator to participate in the program. The evaluator was one who had considerable experience in studies of elementary science teacher training and was aware of the new directions in which elementary science is proceeding. The Principal investigator had considered hiring an elementary science teacher who had been surplused to fill this role, but decided against it because of the belief that a more skilled evaluator would benefit the project more. The evaluator and the Principal investigator discussed the goals of the project, the plan for meeting these goals, and the questions that needed to be explored. Considerable time was devoted to clarifying the kinds of information needed to make sure the training was functioning as intended. The Principal investigator was very interested in studying program implementation as early as possible to make sure everything was

"on track." The evaluator also talked to the trainers to understand their information needs and understand more fully the kinds of data that would help them do their jobs as well as possible. Both the Principal investigator and the trainers expressed a strong interest in knowing the extent to which what was learned was, in fact, being transferred to the classroom.

The evaluator began observing the training sessions on an intermittent basis almost as soon as they started. Although no formal observation system was used, she wrote a brief narrative summary after each observation session detailing the focus of the session, the strategies discussed, and the involvement/engagement level of the participants. After 6 months, she administered a questionnaire to the participants which addressed a wide range of issues, including the adequacy of the training and the extent to which it was used and useful in their own classrooms. She also interviewed the trainers to get their reactions to the project and to hear any new concerns that may have arisen. She had planned to interview other teachers who worked with the participants to assess their awareness of the new teaching techniques (it was hoped that the training would have a "spill-over" effect on others in the school), but this idea was abandoned as being premature after a preliminary review of the teacher questionnaires.

The findings proved to be very useful and the Principal investigator was pleased with the feedback from this early investment. The observational summaries indicated that the training sessions themselves were highly effective. Even during the lecture sessions, participants were engaged and very attentive. The quality of interaction and discussion during the hands-on sessions was very good, with the participants frequently going beyond the demonstration tasks and inventing their own alternatives.

By and large, the interviews with the trainers complemented the observational data. Trainers were pleased at how smoothly their classes were going and at the high level of engagement shown by the participants. They did, however, have some problems with the model teachers and coordinating demonstrations by them with material covered in the other training sessions. Because these teachers were teaching in regular school situations, the needs of the project and the regular classroom too frequently came into conflict.

The data from the participant questionnaire were less positive. While participants had high praise for the training they were receiving, they were somewhat less enthusiastic about the project overall. While they had initially been very pleased with the opportunity to participate, they were finding that the time they spent away from the classroom was interfering with their jobs as teachers. Further, they were unable to apply what they learned to their own classes because they lacked the supplies and materials necessary. The support that had been provided for the traditional science teaching in which they previously had engaged did not meet the needs of the new lessons to which they were being exposed.

Based on these findings, several changes were made in the project. First, video-taped demonstrations were substituted for visits to model teachers during the regular school year. Arrangements were made for demonstrations of model teaching during the summer at a nearby laboratory school that had a summer session. Selection criteria for teachers were also changed so as to require some in-kind support from their institutions. The Principal or someone in the central office had to agree to provide the supplies and materials needed for instructing in the new ways, if such materials were not already available. The Principal investigator also reallocated some of the project funds to provide more materials for the teachers. A number of lessons were designated as ones in which all needed materials would be provided to the participants for use in their classes. Time was also set aside during the training sessions to discuss attempts to apply the new strategies to the classroom. Both successful and unsuccessful applications were considered.

The second year evaluation showed that these changes were having a positive effect. While the second year participants still had a certain level of frustration at being out of their classrooms, their ability to bring back and try out new strategies reduced this frustration greatly. The sharing sessions at which application attempts were discussed became favorites of both the participants and the trainers. The former gained important insights into ways of transferring their skills; the latter gleaned many tips to pass on to the next year's participants.

Example 2: Evaluation of an Integrated Learning System for the Teaching of Mathematics

NSF funded Jones University, along with the Smith School District to conduct a study of the efficacy of the

Boston Integrated Learning System (ILS). The Smith School District, once considered a very fine school system, had over the last several decades, fallen on hard times. Budget cuts, population shifts, and a national economy which has been sliding, combined to give Smith new challenges that it had never before faced. The results were discouraging. Not only were test scores on the decline, but absenteeism was high, and even when in school, the students frequently missed class or were disruptive.

The Boston ILS was selected for implementation in this district because of both its ability to tailor instruction to individual needs and its motivational characteristics. It was also hoped that with an ILS like Boston in place, teachers would be able to spend more one-on-one time with each student, without reducing the quality of instruction received by the group.

This example illustrates the:

- **Problems that can arise when the potential needs of critical stakeholders are not considered**
- **Limitations of relying on a single measure of program impact**
- **Need to provide for formative progress evaluation**
- **Misinterpretations that result from failure to appropriately disaggregate results.**

The Boston ILS provides the hardware and software needed to assist students in elementary mathematics. The ILS combines teaching modules, testing modules, and a reporting component intended to provide an individualized learning experience. It has high quality graphics and an audio component.

Seven schools were selected for the project. The schools were among the most needy in the district, defined in terms of student test scores and free lunch counts. All students in these schools participated.

The goal of the project, as written in the proposal to NSF, was to increase test scores in mathematics. The key indicator of performance was defined as scores on the Schmata Test of Basic Skills, a norm referenced test given every other year to the students.

The study was initiated in the fall of 1989. Because the evaluation design seemed to the Principal Investigator to be very straightforward—pre-post test performance on the Schmata test, no allowance was made for an evaluator at project onset. Rather, the Principal Investigator intended to rely on the normal test reporting of the school district as the means of acquiring evaluative data. He also included some funds for reanalysis of these data by his colleagues at Jones University. The NSF Program Officer queried the Principal Investigator about the scope of the evaluation, raising the question of whether or not it would meet all stakeholders' needs. She asked, specifically, whether or not all relevant parties at the school district had been consulted before

the proposal was submitted. The Principal Investigator said that he had talked to the Director of Curriculum and they were in agreement with regard to the evaluation. He said, however, that he would revisit the questions with a broader group of policymakers at the school level and amend his proposal, if necessary. In the rush of next steps, this consultation fell by the wayside.

Ten months after the project was initiated, the school district faced another budget crisis. The Board asked for evidence that the project was successful, threatening to cut back the in-kind funds that had been allocated for teacher training and planning time to the project schools. The Principal Investigator was able to give his impressions of how the program was working and several teachers also offered their support. In addition, however, the Board received a number of phone calls from other teachers saying that Boston placed too much of a burden on them and the benefit to students was negligible.

Fearing loss of support from the school system, the Principal Investigator called upon some of his colleagues for advice on what to do. Fortunately, the University had on staff some strong educational evaluators. After review of the project and the data available, they came up with an evaluation scheme. Recognizing the fact that the Schmata test results would not be available for more than a year, they gathered some interim measures of program impact. First, they looked at test performance on the assessment modules provided by Boston. Analysis of these data showed that, overall, students were making steady progress and seemed to be retaining the skills learned as measured by the retention tests. (While there was no way to compare these students' performance with that of students given traditional instruction, the analysis at least provided some promise of success.) Second, they looked at data in other areas to see whether an impact could be posited. The areas they selected were attendance and referrals for behavioral incidents. These data showed that, overall, attendance had increased and behavioral incidents had decreased compared to the same time in previous years. Further, when the data for some individual students were examined, these same students showed increased attendance and fewer referrals for disturbance than they had in the past. Taken together, these findings provided weight to the claims of the Principal Investigator and the project continued to receive support from the school district.

Six months later, the evaluators returned to these data and did some additional analyses, disaggregating the results by grade, gender, race, and English-language proficiency. The evaluators found the picture of success which emerged from the overall data did not hold true for each subgroup of students participating. Specifically, they found that both the progress and behavioral data showed striking differences between English and limited-English speaking students, with the limited English speaking students failing to do as well. Despite the fact that the latter were receiving the district's special language supports and were assigned to the regular classroom for their instruction, these students were clearly failing to profit from Boston ILS. Unfortunately, this information was not obtained until after the students had spent a full school year in the project andhad started a second year of participation.

Example 3: The Evaluation Of A Special Program For Gifted Minority Students

Project REACH FOR THE STARS, a project aimed at identifying and supporting gifted minority students in math and science, was funded by NSF under the Comprehensive Regional Center for Minorities Program. This project aimed at identifying talented minority students at the end of 8th grade and encouraging their participation in mathematics and science courses for the duration of their high school career and beyond. It included a mentoring component, Saturday morning and summer enrichment sessions, and support groups.

Students were identified for participation based on test scores, grades, and motivation. Two subgroups were created. Subgroup I, the majority of students (75%), had the highest test scores and at least a "B "average in math and science. The second subgroup, subgroup II (25%), consisted of students who were highly motivated to participate, but had not shown strong performance in the past. Since there were more students who qualified for the program and were interested than those who could be accepted, a lottery system was used to select individuals for participation. Those who were not selected were placed into a comparison group against which to measure the progress of students in subgroup I. Unfortunately, there was not a sufficiently large number of students who fell into subgroup II to allow for a similar procedure.

The evaluation used a wide variety of measures which were entered into a data base student-by-student.

Included were grades both in the target courses, math and science, and in other major academic subjects; test scores from end-of-semester exams; standardized tests; and, as relevant, the SAT, ACT, and College Board results. At the end of the 12th grade, data on post secondary applications and acceptances, as well as scholarships and other honors were obtained and added.

In addition, focus groups were conducted each year with participants in order to get students' reactions to the program both from an academic and a social perspective. Surveys were administered to the parents and teachers at the end of the second and fourth years. Finally, a follow-up survey was sent to graduates (both the participants and the subgroup I comparisons students) one year after they had graduated in order to find out what they were doing, how well they were doing, and what they thought, in retrospect of the special program in which they had participated.

This example illustrates:

- **A Summative Evaluation built on progress data collected annually**
- **The use of both survey and case study methodology**
- **The use of multiple data sources**
- **The problems of interpreting findings without a comparison group**
- **The consideration of timeliness in the production of a report.**

A substudy, which was turned into a dissertation by one part-time researcher, looked closely at the experiences of five students differing in gender, race, and family structure. These case studies were used to provide a thick description of program experiences and student/ family, and staff reactions.

On an annual basis, the data were analyzed for the program participants overall—the comparison group students and for the participating students by subgroups. These annual analysis were used to monitor the progress of the students and to pinpoint individual student problems as they arose. The student focus groups also provided important input for modifications in the project, which fine-tuned the approach.

A final report built on the data collected annually providing an overall summary for the four years of project participation. The only new data added in the fourth year that was not collected previously was the information on honors, awards, and post-secondary acceptances. Because of the careful job that had been done documenting progress along the way, the production of a final report was greatly simplified and there was little protest from any of the participants about the requests for data and the "burden" that it caused.

The final report showed that in general the program was a success. Students made good grades in the

target courses, continued to do well in the other courses, and were accepted into strong post-secondary institutions. There was a statistically significant difference between the grades and test scores of the program participants and the comparison group students. The data collected from the students' parents showed they had high expectations for their children and were convinced by their proven success in the program that they could and should aim high in the future. However, the parents of the nonparticipating students were quite similar in their responses. Where students from subgroup I dropped out or failed for other reasons to succeed, there was usually some extenuating circumstance relating to family or friends. Although the number was too small to be statistically significant overall, there was a tendency for comparison students to drop out more frequently for reasons related to school problems.

Although not all of the subgroup II students were successful, nearly half of them were so. Unfortunately, the analyses did not uncover any particular predictors of who from that group might succeed or fail. Some teachers felt that despite these students failure to attain success in absolute terms, they still performed better than they would have without the program. However, the lack of a comparison group for these students made it impossible to test out this hypothesis. The evaluator felt that understanding of these students could be enhanced by some further interviews or by the data that would result from the follow-up questionnaire and hoped to delay reporting until these data could be analyzed and the picture made more complete. However, the principal investigator felt that the report could not be delayed further without putting in jeopardy any chance of continued funding.

Example 4: Evaluation of a Summer Camp For Female High School Students

Project CAMP CRUSADE FOR WOMEN IN SCIENCE, a five-year project begun in 1987, is aimed at women in high school grades 9–11 and seeks to promote interest and involvement in the study of science. The goals are science-oriented high school and post high school course and activity choices on the part of camp participants which will ultimately lead them to pursue careers in the sciences.

This project, funded by NSF under the Young Scholars Program, currently has an all-woman

staff of two secondary science teachers, two undergraduate students majoring in science, and one college professor.

The project is being carried out by Hill College, a small midwestern institution located in the Mountain school district,a large district with 5 high schools. The participating college professor is the Principal Investigator.

Eligible applicants include all female students in grades 9–11 in the Mountain school district. All applicants are asked to complete a questionnaire which seeks information about previous courses taken, and includes a series of questions measuring attitudes, satisfaction, motivation, educational goals, and career goals. The Principal Investigator considered using Grade Point Average and test scores as selection criteria, but rejected this approach because of questions she had regarding how well they would predict performance on activities at the camp, and because she wanted Camp Crusade to provide opportunities for all women interested in science regardless of their previous attainments.

The summer camp provides opportunities for up to 35 participants to engage in a variety of activities such as lectures, experiments, field studies, films, and study work groups. From among 120 applicants, the 35 participants were randomly selected with equal selections from each grade.

The proposal to NSF included an evaluation component with a modest budget. It called for a Formative Evaluation to be conducted every two years. The project evaluator was a part-time instructor in the Department of Education at the Hill College with prior experience in educational research, but no evaluation experience. The evaluator planned to use an experimental design based on comparisons between the treatment group (camp participants) and a control group which consisted of a random sample of 50 non-accepted applicants. The evaluator attempted to match participants and controls by grade level, but given the small applicant pool this was difficult, since the great majority of applicants were 10th graders.

The first Formative Evaluation of the program took place in 1987. However, the initial evaluation was limited to an Implementation Evaluation that determined that the project was being conducted as planned. No Progress Evaluation was done to determine

whether the participants were moving towards meeting the project's goals. Stakeholders questioned the absence of progress information and wondered whether it was an oversight or an unwillingness to look at the issue.

To meet these concerns, the Principal Investigator and the evaluator decided to modify the evaluation design. They decided that an Implementation Evaluation would be conducted every other year, and a Progress Evaluation would be conducted yearly. The revised evaluation design called for data to be collected both at the end of the summer and during the next school year. Both qualitative and quantitative approaches were to be used to capture a wide variety of information on how the summer experience was affecting the participants. Specifically the data collection would include:

- Questionnaires
- Personal interviews
- Observations
- School records.

During the last two days of the camp, the evaluator conducted personal interviews and observed the work groups; on the last day she distributed a self-administered questionnaire to the participants, The interviews focused primarily on the camp experience while the self-administered questionnaire was essentially a replication of the questionnaire which all applicants had completed. The evaluator had planned to administer this questionnaire also to the control group but could not reach most students in the control group because of the timing of this survey (late summer, before the start of school) and this idea was abandoned.

Analyses of the initial information gathered on the experimental and control students showed that the groups were very similar. There were no differences between the two groups with respect to courses selected or motivation. Overall, the data from the experimental group collected following the camp experience appeared promising. The post-test questionnaire indicated that the camp attendees were more motivated than they had been to pursue advanced and elective high school course work in the sciences, had increased positive attitudes towards science, and were more

likely to consider pursuing future academic studies in the sciences than before they attended the camp. The interviews with the experimental group supported the findings from the questionnaires. They allowed the evaluator to add more qualitative data to the study and enhance the quantitative results.

The observational data suggested that the work group sessions were well-liked by the girls. There was a high level of interaction among the participants, as well as between the participants and the faculty. There was an especially high volume of contact between the campers and the undergraduate counselors. The students were observed asking numerous questions of the counselors regarding their college studies and the nature and difficulty of their programs. There were similar levels of interaction with the high school teachers, but interactions with the professor were constrained. Further observations will be necessary to determine why these interactions were so low, focusing on program design, personalities, etc. This issue was highlighted for further attention in the next Implementation Evaluation.

At the end of the second semester of the school year following the 1988 camp, a review of the school records indicated that 65% of the girls who attended the camp had registered for honors or advanced placement science classes, while only 45% of the control group students had done so. Also, 25% of the experimental group, and only 10% of the control group, chose to take a science-related course as their elective.

A follow-up questionnaire was mailed to both the experimental and control groups six months after the camp. The response rate was low for both groups—50% for the experimental group and 30% for the control group. The Principal Investigator recognized that follow-up phone calls were needed to increase the response rate and insure reliability of findings, but budget constraints prohibited this approach and these data were used in the evaluation although the evaluator was careful to point out their weakness.

The findings indicated that although the measures of attitudes and motivation for the experimental group had decreased slightly since the immediate post-test questionnaire was administered, the measures were significantly higher than those of the control group. Responses to questions about the

This example illustrates:

- **Ability to adapt a design based on new demands**
- **Use of control group for measuring project effects**
- **Utilizing an appropriate mix of data collection techniques - both quantitative and qualitative**
- **Failure to adequately plan for follow-up data collection procedures and correction for low response rates**
- **Over generalizing from one particular sample of participants to a whole population**
- **Treating data from a Progress Evaluation as if they were part of a Summative Evaluation**

camp itself indicated that the participants really enjoyed it and were happy about the new relationships they had established with other students and with the staff. Approximately 25% of the responses indicated that the camp participants desired more environmental science activities. Therefore, the Principal Investigator decided to accommodate this desire by replacing one of the lab activities that received very critical comments with an environmental science activity.

The report on the Progress Evaluation was very positive. Its overall conclusion was that this project had a positive effect and motivated girls to become more involved in science and oriented toward academic studies in scientific fields. In fact, the report included a recommendation that more camps like this should be established throughout the school district and more girls should be encouraged to apply.

The stakeholders who initially called for the Progress Evaluation greeted the findings and recommendation with mixed reactions. Supporters of the project were very enthusiastic about the findings. Critics questioned the conclusions citing the very high initial motivational levels of the experimental and control students and the low response rates in the follow-up survey. They finally agreed that the data were encouraging but that final decisions regarding program expansion needed to await additional data.

CHAPTER SIX: SELECTING AN EVALUATOR

Just as some people presume that anyone can be a teacher or that anyone can be an expert of educational policy, some people believe that anyone can be an evaluator. That is just not the case and any Project Director who works on such an assumption not only endangers the quality of the evaluation effort, but risks being faced with making decisions on poor or incorrect information. The extra time, and sometimes, money it takes to get someone who is properly trained and experienced is well-worth the allocation of resources.

The selection of the appropriate person to be a Project Evaluator is critical to the data collection, monitoring and evaluation tasks that are needed for the management of a project. In an earlier chapter, we brieflytouched on the fact that an evaluator can either be external or internal to the project team. In this section, we talk more about the evaluator focusing on the qualifications that are needed and ways to locate competent personnel.

Skills Needed

The evaluator must be knowledgeable of the project's goals and objectives, and must agree philosophically with the mission and the purpose of the project.

This person must be able to apply evaluation expertise to collect data, monitor and evaluate the project. This person must be able to work independently with minimum supervision and to provide consultative advice, when needed. Further, the evaluator must be able to listen and to deal with all participants and stakeholders in a respectful and sensitive way. There is nothing that can kill an evaluation faster than an arrogant evaluation specialist.

The specific skills needed are listed below:

- Has knowledge of evaluation theory and methodology
- Can differentiate between research and evaluation procedures
- Can plan, design, and conduct an evaluation
- Has knowledge and ability to do statistical analyses of data

- Has knowledge and ability to manage and maintain a moderate- sized database
- Has ability to train staff to input data into the database
- Has skills to write and edit brief interpretive reports
- Has experience conducting an evaluation
- Can communicate clearly and effectively with project and program staff and others related to the evaluation tasks
- Has ability to differentiate between what information is needed solely for the project and what information is needed for the program
- Understands and can do formative evaluation.

Although it is not always possible, it is preferable to have the evaluator as a part of the team from the beginning of the project, even before the project is funded at the proposal stage. This not only benefits the evaluator, but also the entire project. An evaluator who is knowledgeable and sensitive to the kinds of information that will be needed to answer the Formative and Summative questions, can set in place the mechanisms for providing the data needed at project outset. Evaluation will not be seen as an "add-on," as it so often is, and the data gathering can, in all likelihood, be far less intrusive than if it is started at a later date. The fact that evaluators and program implementors may have interests and needs that clash is all too well known. Early establishment of a shared commitment and shared understanding is essential. (In Chapter Five where we present examples of evaluation studies, we illustrate some reasons for having an evaluator on board early on. For more on this issue of the role of the evaluator in new programs and the social conditions that may influence evaluation see Rossi and Freeman, 1993).

Sometimes Project Directors want, or may be pressured, to select as the evaluator someone who has been closely aligned with the project on the program development side—a teacher or curriculum specialist, for example. Assuming the person also has evaluation credentials, such a choice may seem appealing. However, one danger in this route is the strong possibility of a biased evaluation, either real or perceived, taking place. That is, the evaluator may be seen as having too

much ownership in the original idea and, therefore, be unable to conduct an objective evaluation. Indeed, the evaluator herself may feel uncomfortable, torn between possibly conflicting loyalties. Given the stakes that frequently are attached to evaluation findings it is prudent to avoid any apparent conflicts of interest.

Finding an Evaluator

There are many different sources for locating a Project Evaluator. The one that works best will depend on a number of factors including the home institution for the project, the nature of the project, and whether or not the Principal Investigator has some strong feeling about the type(s) of evaluation that are appropriate.

There are at least three avenues that can be pursued:

- If the project is being carried out at or near a college or university, a good starting point is likely to be at the college or university itself. Principal Investigators can contact the Department chairs from areas such as Education, Psychology, Administration, or Sociology and ask about the availability of staff skilled in project evaluation. In most cases, a few calls will yield several names.

- A second source for evaluation assistance comes from independent contractors. There are many highly trained personnel whose major income derives from providing evaluation services. Department chairs may well be cognizant of these individuals and requests to chairs for help might include suggestions for individuals they have worked with outside of the college or university. In addition, independent consultants can be identified from the phone book, from vendor lists kept by procurement offices in state departments of education and in local school systems, and even from resource databases kept by some private foundations, such as the Kellogg Foundation in Michigan.

- Finally, suggestions for evaluators can be obtained from calls to other researchers or perusal of research and evaluation reports. A strong personal recommendation and a discussion of an evaluator's strengths and weaknesses from someone who has worked with a specific evaluator is very useful when starting a new evaluation venture.

Although it may take a chain of telephone calls to get the list started, most Principal Investigators will ultimately find that they have several different sources of evaluation support from which to select. The critical task then becomes negotiating time, content, and, of course, money.

REFERENCES

Rossi, P. H. & Freeman, H. E. (1993). *Evaluation—A Systematic Approach* (5th Edition). Newbury, CA: Sage.

CHAPTER SEVEN: GLOSSARY

Accountability	The responsibility for the justification of expenditures, decisions, or the results of one's own efforts.
Accuracy	The extent to which an evaluation is truthful or valid in what it says about a program, project or material.
Achievement	A manifested performance determined by some type of assessment or testing.
Adversarial/advocacy group	A group of people who enter into cross-examination of counter plans, strategies, or outcomes.
Affective	Consists of emotions, feelings, and attitudes.
Algorithm	A step-by-step problem-solving procedure.
Anonymity (provision for)	Evaluator action to ensure that the identity of subjects cannot be ascertained during the course of a study, in study reports, or in any other way.
Assessment	Often used as a synonym for evaluation. The term is sometimes recommended for restriction to processes that are focused on quantitative and/or testing approaches.
Attitude	A person's mental set toward another person, thing, or state.
Attrition	Loss of subjects from the defined sample during the course of a longitudinal study.
Audience(s)	Consumers of the evaluation; those who will or should read or hear of the evaluation, either during or at the end of the evaluation process. Includes those persons who will be guided by the evaluation in making decisions and all others who have a stake in the evaluation (see stakeholders).
Background	The contextual information that describes the reasons for the project, its goals, objectives, and stakeholders' information needs.
Baseline	Facts about the condition or performance of subjects prior to treatment or intervention.

Behavioral objectives	Specifically stated terms of attainment to be checked by observation, or test/measurement.
Bias	A consistent alignment with one point of view.
Case Study	An intensive, detailed description and analysis of a single project, program, or instructional material in the context of its environment.
Checklist approach	Checklists are the principal instrument for practical evaluation; especially for investigating the thoroughness of implementation.
Client	The person or group or agency that commissioned the evaluation.
Coding	To translate a given set of data or items into machine-readable categories
Cognitive	The domain of knowledge—"knowledge-that" or "knowledge-how."
Cohort	A term used to designate one group among many in a study. For example, "the first cohort" may be the first group to have participated in a training program.
Comparison group	A group that provides a basis for contrast with (in experimentation) an experimental group (i.e., the group of people participating in the program or project being evaluated). The comparison group is not subjected to the treatment (independent variable), thus creating a means for comparison with the experimental group that does receive the treatment. Comparison groups should be "comparable" to the treatment group, but can be used when close matching is not possible (see also Control Group).
Component	A physically or temporally discrete part of a whole. It is any segment that can be combined with others to make a whole.
Conceptual scheme	A set of concepts that generate hypotheses and simplify description.
Conclusions (of an evaluation)	Final judgments and recommendations.
Content analysis	A process of systematically determining the characteristics of a body of material or practices.

Control group	A group that does not receive the treatment (service or product). The function of the control group is to determine the extent to which the same effect occurs without the treatment. The control group must be closely matched to the experimental group.
Correlation	A statistical measure of the degree of relationship between variables.
Cost analysis	The practical process of calculating the cost of something that is being evaluated. Cost analysis looks at: (1) costs to whom; (2) costs of what type; and (3) costs during what period.
Cost-benefit analysis	This process estimates the overall cost and benefit of each alternative product or program.
Cost-effectiveness	This analysis determines what a program or procedure costs against what it does (effectiveness). Is this product or program worth its costs?
Criterion, criteria	A criterion (variable) is whatever is used to measure as success, e.g., grade point average.
Criterion-referenced test	Tests whose scores are interpreted by referral to well defined domains of content or behaviors, rather than by referral to the performance of some comparable group of people.
Cross-sectional study	A cross-section is a random sample of a population, and a cross-sectional study examines this sample at one point in time. Successive cross-sectional studies can be used as a substitute for a longitudinal study. For example, examining today's first year students and today's graduating seniors may enable the evaluator to infer that the college experience has produced or can be expected to accompany the difference between them. The cross sectional study substitutes today's seniors for a population that cannot be studied until 4 years later.
Delivery system	The link between the product or service and the immediate consumer (the recipient population).
Dependent variable	One that represents the outcome—the contrast is with independent variables some of which can be manipulated.
Descriptive statistics	Those that involve summarizing, tabulating, organizing, and graphing data for the purpose of describing objects or individuals that have been measured or observed.

Design	The process of stipulating the investigatory procedures to be followed in doing a certain evaluation.
Dissemination	The process of communicating information to specific audiences for the purpose of extending knowledge and, in some cases, with a view to modifying policies and practices.
Effectiveness	Refers to the conclusion of a Goal Achievement Evaluation. "Success" is its rough equivalent.
Executive report	An abbreviated report that has been tailored specifically to address the concerns and questions of a person whose function is to administer an educational program or project.
Executive summary	A nontechnical summary statement designed to provide a quick overview of the full-length report on which it is based.
Experimental design	The plan of an experiment, including selection of subjects who receive treatment and control group (if applicable), procedures, and statistical analyses to be performed.
Experimental group	The group that is receiving the treatment.
External evaluation	Evaluation conducted by an evaluator from outside the organization within which the object of the study is housed.
Extrapolate	To infer an unknown from something that is known. (Statistical definition—to estimate the value of a variable outside its observed range.)
False positive	When an event is predicted and it does not occur (Type I error).
False negative	When an event is not predicted and it occurs (Type II error).
Feasibility	The extent to which an evaluation is appropriate for implementation in practical settings.
Field test	The study of a program, project, or instructional material in settings like those where it is to be used. Field tests may range from preliminary primitive investigations to full-scale summative studies.

Flow chart	A graphic representation of a set of decisions that is set up to guide the management of projects, including evaluation projects.
Focus group	A group selected for its relevance to an evaluation that is engaged by a trained facilitator in a series of discussions designed for sharing insights, ideas, and observations on a topic of concern.
Formative evaluation	Evaluation designed and used to improve an intervention, especially when it is still being developed.
Gain scores	The difference between a student's performance on a test and his or her performance on a previous administration of the same or parallel test.
Generalizability	The extent to which information about a program, project, or instructional material collected in one setting can be used to reach a valid judgment about how it will perform in other settings.
Goal-free evaluation	Evaluation of outcomes in which the evaluator functions without knowledge of the purposes or goals.
Hawthorne effect	The tendency of a person or group being investigated to perform better (or worse) than they would in the absence of the investigation, thus making it difficult to identify treatment effects.
Hypothesis testing	The standard model of the classical approach to scientific research in which a hypothesis is formulated before the experiment to test its truth. The results are stated in probability terms that the results were due solely to chance. The significance level of one chance in 20 (.05) or one chance in 100 (.01) is a high degree of improbability.
Impact evaluation	An evaluation focused on outcomes or pay-off.
Implementation evaluation	Assessing program delivery (a subset of Formative Evaluation).
Indicator	A factor, variable, or observation that is empirically connected with the criterion variable, a correlate. For example, judgment by students that a course has been valuable to them for pre-professional training is an indicator of that value.
Inferential statistics	These statistics are inferred from characteristics of samples to characteristics of the population from which the sample comes.

Informed consent Agreement by the participants in an evaluation of the use of their names and/or confidential information supplied by them in specified ways, for stated purposes, and in light of possible consequences prior to the collection and/or release of this information in evaluation reports.

Instrument An assessment device (test, questionnaire, protocol, etc.) adopted, adapted, or constructed for the purpose of the evaluation.

Interaction Two factors or variables interact if the effect of one, on the phenomenon being studied, depends upon the magnitude of the other. For example, mathematics education interacts with age, being more or less effective depending upon the age of the child.

Internal evaluator Internal evaluations are those done by project staff, even if they are special evaluation staff, that is, external to the production/writing/ teaching/service part of the project.

Level of significance The probability that the observed difference occurred by chance.

Longitudinal study An investigation or study in which a particular individual or group of individuals is followed over a substantial period of time to discover changes due to the influence of the treatment, or maturation, or environment.

Mastery level The level of performance needed on a criterion. The mastery level is often arbitrary.

Matching An experimental procedure in which the subjects are so divided, by means other than lottery, that the groups are regarded for the purposes at hand to be of equal merit or ability. (Often matched groups are created by ensuring that they are the same or nearly so on such variables as sex, age, grade point averages, and past test scores.)

Matrix An arrangement of rows and columns used to display components of evaluation design.

Mean Also called "average" or arithmetic average. For a collection of raw test scores, the mean score is obtained by adding all scores and dividing by the number of people taking the test.

Measurement	Determination of the magnitude of a quantity.
Median	The point in a distribution which divides the group into two, as nearly as possible. For example, in a score distribution, half the scores fall above the median and half fall below.
Meta-analysis	The name for a particular approach to synthesizing quantitative studies on a common topic, involving the calibration of a specific parameter for each ("effect size").
Metric data	Data which includes a unit of measurement (i.e., dollars, inches).
Mode	The value which occurs more often than any other. If all scores (in a score distribution) occur with the same frequency, there is no mode. If the two highest score values occur with the same frequency, there are two modes.
Needs assessment	Using a diagnostic definition, need is anything essential for a satisfactory mode of existence or level of performance. The essential point of a needs assessment for evaluation is the identification of performance needs.
Nominal data	Data which consist of categories only without order to these categories (i.e., region of the country, courses offered by an instructional program).
"No significant difference"	A decision that an observed difference between two statistics occurred by chance.
Nonreactive measures	Assessments done without the awareness of those being assessed.
Norm	A single value, or a distribution of values, constituting the typical performance of a given group.
Norm-referenced tests	Tests that measure the *relative* performance of the individual or group by comparison with the performance of other individuals or groups taking the same test.
Objective	A specific description of an intended outcome.
Observation	The process of direct sensory inspection involving trained observers.
Operational definition	A definition of a term or object achieved by stating the operations or procedures employed to distinguish it from others.

Ordered data	Non-numeric data in ordered categories (for example, students performance being categorized as excellent, good, adequate, and poor).
Outcome	Post-treatment or post-intervention effects.
Paradigm	A general conception of or model for a discipline or subdiscipline which may be very influential in shaping its development. (For example, "The classical social science paradigm in evaluation.")
Peer review	Evaluation done by a panel of judges with qualifications approximating those of the author or candidate.
Performance-based	The use of global ratings of behavior assessment which is a movement away from paper-and-pencil testing. This assessment is costly and there may be a loss of validity and reliability.
Pilot test	A brief and simplified preliminary study designed to try out methods to learn whether a proposed project or program seems likely to yield valuable results.
Planning evaluation	Evaluation planning is necessary before a program begins, both to get baseline data, and to evaluate the program plan, at least for evaluability. Planning avoids designing a program that is unevaluable.
Population	All persons in a particular group.
Post test	A test to determine performance after the administration of a program, project, or instructional material.
Pretest	A test to determine performance prior to the administration of a program, project, or instructional material. Pretests serve two purposes: diagnostic and baseline. Also the use of an instrument (questionnaire, test, observation schedule) with a small group to detect need for revisions.
Process evaluation	Refers to the evaluation of the treatment or intervention. It focuses entirely on the variables between input and output.
Product	A pedagogical process or material coming from research and development.
Program	The general effort that marshals staff and projects toward defined and funded goals.
Progress evaluation	A subset of Formative Evaluation.

Prompt	Reminders used by interviewers to obtain complete answers.
Qualitative evaluation	The part of the evaluation that is primarily descriptive and interpretative, and may or may not lend itself to quantitative treatment.
Quantitative evaluation	An approach involving the use of numerical measurement and data analysis based on statistical methods.
Quasi-experimental	When a random allocation of subjects to experimental and control groups can not be done, a quasi-experimental design can seek to simulate a true experimental design, by identifying a group that closely matches the experimental group.
Random	Affected by chance.
Random sampling	Drawing a number of items of any sort from a larger group or population so that every individual item has a specified probability of being chosen.
Recommendations	Suggestions for specific appropriate actions based upon analytic approaches to the program components.
Reliability	Statistical reliability is the consistency of the readings from a scientific instrument or human judge.
Remediation	The process of improvement or a recommendation for a course of action or treatment that will result in improvement.
Replication	Repeating an intervention or evaluation with all essentials unchanged. Replications are often difficult to evaluate because of changes in design or execution.
Research	The general field of disciplined investigation.
Response bias	Error due to incorrect answers.
Sample	A part of a population.
Sample bias	Error due to non-response or incomplete response from selected sample subjects.
Sampling error	Error due to using a sample instead of entire population from which sample is drawn.
Secondary data analysis	A reanalysis of data using the same or other appropriate procedures to verify the accuracy of the results of the initial analysis or for answering different questions.

Self-administered instrument	A questionnaire or report completed by a study participant without the assistance of an interviewer.
Self-report instrument	A device in which persons make and report judgments about the functioning of their project, program, or instructional material.
Significance	Overall significance represents the total *synthesis* of all you have learned about the merit or worth of the program or project. This is different from statistical significance which may be testing one of several conditions of a program or project.
Stakeholder	A program's stakeholder is one who has credibility, power, or other capital invested in the project, and thus can be held to be to some degree at risk with it.
Standard deviation	A measure of the spread of a variable, based on deviation from the mean value for metric data.
Standardized tests	Tests that have standardized instructions for administration, use, scoring, and interpretation with standard printed forms and content. They are usually norm-referenced tests but can also be criterion-referenced.
Statistic	A summary number that is typically used to describe a characteristic of a sample.
Strategy	A systematic plan of action to reach predefined goals.
Summary	A short restatement of the main points of a report.
Summative evaluation	Evaluation designed to present conclusions about the merit or worth of an intervention and recommendations about whether it should be retained, altered, or eliminated.
Time series study	A study in which periodic measurements are obtained prior to, during, and following the introduction of an intervention or treatment in order to reach conclusions about the effect of the intervention.
Treatment	Whatever is being investigated; in particular, whatever is being applied or supplied to, or done by, the experimental groups that is intended to distinguish them from the comparison groups.
Triangulation	In an evaluation, it is an attempt to get a fix on a phenomenon or measurement by approaching it via several independent routes. It can be more than three routes. This effort provides redundant measurement.

Unanticipated outcomes A result of a program or interview that was unexpected. Often used as a synonym for side-effects, but only a loose equivalent.

Utility The extent to which an evaluation produces and disseminates reports that inform relevant audiences and have beneficial impact on their work.

Utilization (of evaluations) Use and impact are terms used as substitutes for utilization. Sometimes seen as the equivalent of implementation, but this applies only to evaluations which contain recommendations.

Validity The soundness of the use and interpretation of a measure.

SOURCES

Jaeger, R. M. (1990). *Statistics A Spectator Sport*. Newbury Park, CA: Sage.

Joint Committee on Standards for Educational Evaluations (1981). *Standards for Evaluation of Educational Programs, Projects and Materials*. New York, NY: McGraw Hill.

Rogers, E. M. (1983). *Diffusion of Innovations*. New York, NY: Free Press.

Scriven, M. (1991). *Evaluation Thesaurus*. Fourth Edition. Newbury Park, CA: Sage.

Authors of Chapters 1-4.

CHAPTER EIGHT: ANNOTATED BIBLIOGRAPHY

In selecting publications for inclusion in this short bibliography, an effort was made to incorporate those which can be useful for evaluators who want to find information relevant to the tasks they will be faced with and which this brief handbook could not cover in depth. Thus, we have not necessarily included all books which might be on a reading list in a graduate course dealing with educational evaluation, but have selected those which NSF/HRD grantees should find most useful. This includes most of those referenced in Chapters 1-4, as well as others which provide perspectives and methods not extensively covered in the handbook.

We have divided this bibliography into two sections. Section One includes books that deal in large part with the broad topics of theory and practice in the field of evaluation, with emphasis on those which deal specifically with education. Section Two is more narrowly focused on technical topics and/or single issues. Of course, this is not an air-tight division: many of the broad-based works contain a great deal of technical information and hands-on advice.

Section One: Theory and Practice

House, Ernest R. (1993). *Professional Evaluation — Social Impact and Political Consequences.* Newbury Park, CA: Sage.

The author is a professor in the School of Education at the University of Colorado and an experienced evaluator of major social and educational programs. In the author's own words, this book: "is about analyzing the social, political, economic, historical, and cultural influences on evaluation." It provides a thoughtful overview of the field's evolution, from its earlier reliance on "value-free" experimental and primarily quantitative methods to the current emphasis on methods more appropriate to the diverse, multicultural and politically charged environment in which social and educational programs operate.

Rossi, Peter H. & Freeman, Howard E. (1993). *Evaluation — A Systematic Approach* (5th Edition). Newbury Park, CA: Sage.

This is the newest edition of one of the most comprehensive and widely used texts about evaluation. It provides extensive and sophisticated discussions of all aspects of designing and assessing the implementation and utility of social programs. Many of the projects cited and discussed in this volume deal with educational programs and innovations, although the bulk of the programs with which these authors had first hand experience were in the fields of housing, health services, and criminal justice. Most tasks that evaluators are likely to be asked to perform, and most problems they will have to deal with, technical as well as political, are covered here. The authors adhere to the social science model in their approach to evaluation, with clear preference for randomized and quasi-experimental designs, but they also cover other evaluation methods, including the use of qualitative and judgmental approaches.

Worthen, Blaine R. & Sanders, James R. (1987). *Educational Evaluation: Alternative Approaches and Practical Guidelines.* White Plains, NY: Longman.

This book was designed primarily as a basic text for graduate courses in evaluation, or related administration, curriculum, or teacher education courses, where efforts are made to teach practitioners how to assess the effectiveness of their educational endeavors. It seeks to familiarize readers with alternative approaches for planning evaluations, and provides step-by-step practical guidelines for conducting them. The book is very systematically organized, describes at length all evaluation approaches which have been developed over the years, and includes a great deal of useful information, especially for the inexperienced evaluator. A detailed, naturalistic description of the conduct of an evaluation program, including problems encountered with school staff, other stakeholders, and administrators, provides a useful example of "real world" issues in evaluation.

Scriven, Michael (1991). *Evaluation Thesaurus* (4th Edition). Newbury Park, CA: Sage.

A highly original, wide ranging collection of ideas, concepts, positions and techniques which reflect the critical, incisive, and often unconventional views held by this leader in the field of evaluation. It contains a 40-page introductory essay on the nature of evaluation andnearly 1,000 entries which range from one paragraph definitions of technical terms and acronyms to philosophical and methodological discussions extending over many pages. The Thesaurus is not focused on the field of education, but it provides excellent coverage of issues and concepts of interest to educational evaluators.

Guba, E. G. & Lincoln, Y. S. (1989). *Fourth Generation Evaluation.* Newbury Park, CA: Sage.

The authors propose a monumental shift in evaluation practice, advocating the constructionist position in its most extreme form. Guba and Lincoln describe problems faced by previous generations of evaluators—politics, ethical dilemmas, imperfections and gaps, inclusive deductions—and lay the blame for failure and nonutilization at the feet of the unquestioned reliance on the scientific/positivist paradigm of research. Fourth generation evaluation moves beyond science to include the myriad human, political, social, cultural, and contextual elements that are involved in the evaluation process.

The book describes the differences between the earlier (and still widely used) evaluation model, based on positivist/scientific assumptions and statistical techniques, and the "naturalistic" approach to evaluation, and outlines methodological guidelines for the conduct of naturalistic evaluations.

Joint Committee on Standards for Educational Evaluation (1981). *Standards for Evaluations of Educational Programs, Projects, and Materials.* New York, NY: McGraw Hill.

For more than 10 years this volume has provided guidance to educational evaluators. It contains 30 standards for program evaluation along with examples of their application. It is the authoritative reference on principles of good and ethical program evaluation. Sponsored by several professional associations concerned with the quality of program evaluations in education, the Joint Committee has defined state-of-the-art principles, guidelines, and illustrative cases that can be used to judge the quality of any evaluation. The standards fall into four categories: utility, feasibility, propriety, and accuracy.

Patton, Michael Q. (1986). *Utilization-Focused Evaluation* (2nd Edition). Newbury Park, CA: Sage.

In a book that combines both the theoretical and the practical, Patton examines how and why to conduct evaluations. In this revised and updated edition, the author provides practical advice grounded in evaluation theory and practice, and shows how to conduct evaluations from beginning to end in ways that will be useful—and used. This volume discusses the ferment and changes in evaluation during the eighties and the tremendous growth of "in-house" evaluations conducted by internal evaluators. Patton also discusses a methodological synthesis of the "qualitative versus the quantitative" methods debate, as well as the cross-cultural development of evaluation as an internationally recognized profession. Presented in an integrated framework, the author discusses topics such as utilization built on a new, concise definition of evaluation which emphasizes providing useful and usable information to specific people.

Section Two: Technical Topics

Jaeger, R. M. (1990). *Statistics: A Spectator Sport* (2nd Edition). Newbury Park, CA: Sage.

This book takes the reader to the point of understanding advanced statistics without introducing complex formulas or equations. It covers most of the statistical concepts and techniques which evaluators commonly use in the design and analysis of evaluation studies, and most of the examples and illustrations are from actual studies performed in the field of education. The topics included range from descriptive statistics, including measures of central tendency and fundamentals of measurement, to inferential statistics and advanced analytic methods.

Campbell, Donald T. & Stanley, Julian C. (1966). *Experimental and Quasi-experimental Designs for Research.* Boston, MA: Houghton Mifflin.

This slim (84 pages) volume is a slightly enlarged version of the chapter originally published in the 1963 ***Handbook of Research on Teaching*** and is considered the classic text on valid experimental and quasi-experimental designs in real world situations where the experimenter has very limited control over the environment. To this day, it is the most useful basic reference book for evaluators who plan the use of such designs.

Herman, Joan L. (Editor, 1987). *Program Evaluation Kit.* Newbury Park, CA: Sage.

This kit, prepared by the Center for the Study of Evaluation at the University of California, Los Angeles, contains nine books written to guide and assist evaluators in planning and executing evaluations, with emphasis on practical, field-tested step-by-step procedures and with considerable attention to the management of each phase. The kit makes heavy use of charts, illustrations, and examples to clarify the material for novice evaluators. Volume 1, ***Evaluator's Handbook***, provides an overview of evaluation activities, describes the evaluation perspective which guides the kit, and discusses specific procedures for conducting Formative and Summative evaluations. The remaining eight volumes deal with specific topics:

Volume Two ***How to Focus an Evaluation***

Volume Three ***How to Design a Program Evaluation***

Volume Four ***How to Use Quantitative Methods in Evaluation***

Volume Five ***How to Assess Program Implementation***

Volume Six ***How to Measure Attitudes***

Volume Seven ***How to Measure Performance and Use Tests***

Volume Eight ***Hot to Analyze Data***

Volume Nine ***How to Communicate Evaluation Findings***

Depending on their needs, evaluators will find every one of these volumes useful. Volume Seven, ***How to Measure Performance and Use Tests***, covers a topic for which we have not located another suitable text for inclusion in this bibliography.

The Kit can be purchased as a unit, or by ordering individual volumes.

Yin, Robert K. (1989). Case Study Research: Design and Method. Newbury Park, CA: Sage.

The author's background in experimental psychology may explain the emphasis in this book on the use of rigorous methods in the conduct and analysis of case studies, thus minimizing what many believe is a spurious distinction between quantitative and qualitative studies. While arguing eloquently that case studies are an important tool when an investigator (or evaluator) has little control over events and when the focus is on a contemporary phenomenon within some real-life context, the author insists that case studies be designed and analyzed so as to provide generalizable findings. Although the focus is on design and analysis, data collection and report writing are also covered.

Fowler, Floyd J., Jr. (1993). *Survey Research Methods* (2nd Edition). Newbury Park, CA: Sage.

Using non-technical language, the author has provided a comprehensive discussion of survey design (including sampling, data collection methods, and the design of survey questions) and procedures which constitute good survey practice, including attention to data quality and ethical issues. According to the author, "this book is intended to provide perspective and understanding to those who would be designers or users of survey research, at the same time as it provides a sound first step for those who actually may go about collecting data."

Stewart, David W. & Shamdasani, Prem N. (1990). *Focus Groups: Theory and Practice.* Newbury Park, CA: Sage.

This book differs from many others published in recent years which address primarily techniques of recruiting participants and the actual conduct of focus group sessions. This volume pays considerable attention to the fact that focus groups are by definition an exercise in group dynamics, which must be taken into account when conducting such groups and especially when interpreting the results obtained. However, it also covers very adequately practical issues such as recruitment of participants, the role of the moderator, and appropriate techniques for data analysis.

Linn, Robert L., Baker, Eva L., & Dunbar, Stephen B. (1991). Complex, performance-based assessment: expectations and validation criteria. *Educational Researcher*, (Vol. 20, No. 8, pp. 15-22).

In recent years there has been an increasing emphasis on assessment results, as well as increasing concern about the nature of the most widely used forms of student assessment and uses that are made of the results. These conflicting forces have helped create a burgeoning interest in alternative forms of assessments, particularly complex, performance-based assessments. It is argued that there is a need to rethink the criteria by which the quality of educational assessments are judged, and a set of criteria that are sensitive to some of the expectations for performance-based assessments is proposed.